Research Article

A generic statistical approach for modelling error of geometric features in GIS

WENZHONG SHI
Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, PR China
email: lswzshi@polyu.edu.hk

(Received 28 January 1997; accepted 10 June 1997)

Abstract. This paper describes a newly developed statistical approach for modelling positional error of geometric features in GIS. The generic statistical models for N-dimensional features are firstly derived. The models for one- and two-dimensional features are then developed as the specific cases of the generic models. In each dimension, the GIS features are classified as points, line segments and line features. Because of the errors, features stored in GIS may not correspond with their actual location in the real world. The true location of a GIS feature is only known within a certain area around the represented location in GIS. This newly developed approach can be used to provide a statistical description of such areas. For one-, two- and N-dimensional GIS features, they are defined as confidence intervals, confidence regions and confidence spaces respectively. The areas are related to the positional errors of the composite points of the features and to the predefined confidence level. The models are derived based on the assumptions that the errors of the composite points are independent and follow multi-dimensional normal distributions.

1. Introduction

Error of GIS features is one of the critical issues in today’s GIS research society. Placing unrealistically high trust in the accuracy of data in GIS may cause loss for GIS users. For instance, the positional error of a cadastral boundary in a GIS may cause error in land parcel area calculation, which may lead to the loss of land for the owner. This loss can be serious, especially in an area of high land price, such as Hong Kong. Errors in GIS data present two major problems: few GIS can provide users with error indicators for the data, and; the nature of errors in GIS data is not theoretically clear. This paper will deal with the second problem. There are several aspects of errors in GIS: positional, attribute, topological and temporal errors. This paper concentrates upon positional error of GIS features.

Normally, the true location of points is not exactly at the position represented in a GIS, since there are errors inherited from measurement of the points. The true location of a point is actually within an area somewhere around its measured location in the GIS. In this paper, a statistical approach to modelling positional errors of GIS features is presented. This approach provides a method of describing the nature (its geometric location and shape) of such an area around a measured point, line segment or line.

Interest in the accuracy of points has a long history in surveying and mapping, and there are a number of models which have been developed for describing the
error of surveyed points (Mikhail and Ackermann 1976). The positional error of a two-dimensional line is more important for digital cartography and GIS than traditional surveying. Research on positional error of a two-dimensional line segment or line can be classified into three categories: (a) error of points on a line segment, (b) error distribution of the line, and (c) confidence region.

The positional error of a point on a line segment can be derived from the law of error propagation that is extensively used in surveying. Many researchers have worked on this problem, such as Caspary and Scheuring (1992) and Shi (1994), concluding that the error of the middle point coordinates is smaller than that of the two composite endpoints of the line segment. The error of the coordinates at the midpoint is 0.7 times that of the two endpoints, when endpoints are independent and have the same accuracy.

To find out the error distribution of line segments and lines, Dutton (1992) simulated the distribution of line segments using Monte Carlo simulation. Shi (1994), on the other hand, derived the error distribution of a line segment using statistics and probability theory. In that work, the error distribution of the line segment was represented by a set of one-dimensional distributions and two two-dimensional normal distributions. The error distribution is very important for spatial analysis, and is based on the assumption that we have multiple measurements for each feature in the GIS. However, there is only one measurement for each feature in most of the current, operational GIS. The error distribution derived is thus mainly useful theoretically at this stage.

Given the fact of a single measurement of each geometric feature in a GIS, Shi (1994) derived the concept of a confidence region for a two-dimensional line segment using a rigorous statistical derivation. For any line this method can be used to derive a region around the measured line segment which includes the true location of the line segment. In the model, the relationship of the confidence region with the error of the composite points and the predefined confidence level is clearly defined. The shape of the confidence region is different from that of the epsilon band and closer to reality. The epsilon band (Perkal 1966) also defines a confidence region, on the basis of a constant distance from either side of the line and from its two endpoints. It can be described as the area occupied by rolling a ball along the line. The epsilon model, however, does not define the relationship of the band width with confidence level. Chrisman (1982) further developed the epsilon band model with practical measurements and applied them for modelling error of a GIRAS digital file, the US Geological Surveys Land Use/Land Cover data.

All the above discussions on the positional error of geometric features in GIS refer to two-dimensional space. GIS have also been applied to three-dimensional problems, such as geological applications, and to four-dimensional applications, such as spatio-temporal ones. Accordingly, we have to cope with error within each of these dimensions. It is important for both theoretical studies and applications of GIS to develop models that handle positional errors for N-dimensional GIS features. With a generic error model for N-dimensional GIS, such as that presented here, we can resolve more general problems for four, five and higher dimensional GIS when they come.

2. Geometric features in GIS

A geometric feature in a GIS may be one-dimensional, such as a road section or a river reach, which might be used in dynamic segmentation for attributing two-dimensional linear features. Geometric features in two-dimensional space are the most popular in current GIS applications. Three-dimensional GIS are required in
A generic statistical approach for modelling error in GIS

<table>
<thead>
<tr>
<th></th>
<th>Point</th>
<th>Line Segment</th>
<th>Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-D</td>
<td><img src="image" alt="1D features" /></td>
<td><img src="image" alt="1D features" /></td>
<td><img src="image" alt="1D features" /></td>
</tr>
<tr>
<td>2-D</td>
<td><img src="image" alt="2D features" /></td>
<td><img src="image" alt="2D features" /></td>
<td><img src="image" alt="2D features" /></td>
</tr>
</tbody>
</table>

Figure 1. One- and two-dimensional geometric features in GIS. In each dimension, the features are classified as point, line segment and line.

In many application fields, such as geology. The spatio-temporal GIS deals with four-dimensional GIS features if the spatial one is three-dimensional. Therefore, we introduce the concept of error in the N-dimensional space of GIS to generalize the discussions in this paper.

In each space, the geometric features of GIS can be classified as point, line segment or line. In describing positional error of geometric features, the polygon type can be considered as a special line—a closed line. Therefore, the scope of this paper is identified as: point, line segment and line features in one-, two- and N-dimensional spaces. The geometric features in one- and two-dimensional GIS are illustrated in figure 1.

The next step is to model positional error of the geometric features in the one-, two- and N-dimensional spaces. The error of points in a GIS is determined by the error of original data (e.g. a hard copy map), data capture processes (e.g. digitizing) and spatial operations in the GIS. The error of points in the GIS follows a normal distribution if all these error sources are independent and equally contribute to the error of the points (Shi 1994). Error in the GIS can be either independent or correlated. In this study, we consider errors between points to be statistically independent. When the errors are correlated to each other, they can be modelled by using stochastic techniques. The nature of the error of an arbitrary point on a line segment under the condition that errors are independent is different from the correlated case. Detailed discussions on error models for correlated cases are in Shi and Liu (1997).

In the following, we first derive the error models for N-dimensional geometric features in GIS. These are the generic models. The error models for one- and two-dimensional geometric features can be further developed from the generic models. These models are the specific cases of the generic models.

### 3. Positional error of N-dimensional features

#### 3.1. An N-dimensional point

In N-dimensional space, a point is defined as a measured N-dimensional stochastic point $Q_{x0} = [X_{10}, X_{20}, ..., X_{N0}]^T$ that follows an N-dimensional normal
distribution \((N_N)\) as:

\[
Q_{N0} = \begin{bmatrix}
X_{10} \\
X_{20} \\
\vdots \\
X_{N0}
\end{bmatrix} \sim N_N
\begin{bmatrix}
\mu_{10} \\
\mu_{20} \\
\vdots \\
\mu_{N0}
\end{bmatrix}
\begin{bmatrix}
\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\
\sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2
\end{bmatrix}
\]

where \(\mu, \sigma^2, \sigma_{ij} = \sigma_{ji} (i, j = 1, 2, \ldots, N)\) are the parameters that define the statistical properties of the stochastic vector. The measured point \(Q_{N0}\) is an estimator of the true value of the point \(\phi_{N0} = [\mu_{10}, \mu_{20}, \ldots, \mu_{N0}]^T\).

3.1.1. The confidence space of an \(N\)-dimensional point

To describe the error of the measured vector in \(N\)-dimensional space, we derive a sub-space, named as the confidence space, around the measured \(N\)-dimensional vector \(Q_{N0}\) that contains the true location of the point. The confidence space \(J_{N0}\) of the vector is such that the measured point \(\phi_{N0}\) is contained within the space \(J_{N0}\) with a probability that is larger than a prescribed confidence level \(\gamma\), i.e.

\[
P(\phi_{N0} \in J_{N0}) \geq \gamma
\]

The \(J_{N0}\) is a set \([x_1, x_2, \ldots, x_N]^T\) with \(x_i\) satisfying

\[
X_\theta - a_{Ni} \leq x_i \leq X_\theta + a_{Ni}, \ i = 1, \ldots, N
\]

where

\[
a_{Ni} = k^{1/2} \sigma_i
\]

\(k = \chi^2_{N-1, 1-\gamma/N}\). The parameter \(k\) is dependent on the selected confidence level \(\gamma\) and can be looked up in a chi-square distribution table. With a risk expressed by the confidence level we can state that the true point is somewhere inside the confidence space.

3.2. \(N\)-dimensional line segments

Two measured points defining an \(N\)-dimensional line segment have stochastic vectors \(Q_{N1} = [X_{11}, X_{21}, \ldots, X_{n1}]^T\) and \(Q_{N2} = [X_{12}, X_{22}, \ldots, X_{n2}]^T\). They follow the \(N\)-dimensional normal distributions \((N_N)\):

\[
Q_{N1} = \begin{bmatrix}
X_{11} \\
X_{21} \\
\vdots \\
X_{N1}
\end{bmatrix} \sim N_N
\begin{bmatrix}
\mu_{11} \\
\mu_{21} \\
\vdots \\
\mu_{N1}
\end{bmatrix}
\begin{bmatrix}
\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\
\sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2
\end{bmatrix}
\]

\[
Q_{N2} = \begin{bmatrix}
X_{12} \\
X_{22} \\
\vdots \\
X_{N2}
\end{bmatrix} \sim N_N
\begin{bmatrix}
\mu_{12} \\
\mu_{22} \\
\vdots \\
\mu_{N2}
\end{bmatrix}
\begin{bmatrix}
\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\
\sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2
\end{bmatrix}
\]

where \(\mu_{il}, \mu_{l2}, \sigma_i^2, \sigma_{ij} = \sigma_{ji} (i, j = 1, \ldots, N)\) are the parameters that define the statistical
properties of two endpoints. The points $Q_{N1}$ and $Q_{N2}$ are estimators of the true endpoints $\phi_{N1} = [\mu_{11}, \mu_{21}, \ldots, \mu_{N1}]^T$ and $\phi_{N2} = [\mu_{12}, \mu_{22}, \ldots, \mu_{N2}]^T$ of the line segment.

An N-dimensional line segment is determined by a set of points on it. An arbitrary point on the true line segment between $w_{N1}$ and $w_{N2}$ is represented by

$$w_{Nr} = C_{m1}r_m_{21} \ldots m_{Nr}D_{r} = (1 - r)w_{N1} + rw_{N2} = (1 - r) + r \quad r \in [0, 1]$$

3.2.1. The distribution of points on an N-dimensional line segment

The parameters of the distribution of an arbitrary point on the N-dimensional line segment can be derived from those of the points $Q_{N1}$ and $Q_{N2}$ based on the law of statistical propagation (Serfling 1980). According to equation (7), an arbitrary point on a measured line segment can be represented by

$$Q_{Nr} = X_{1r} \ldots X_{Nr} = (1 - r)Q_{N1} + rQ_{N2} = (1 - r) + rX_{11} \ldots X_{N2}$$

Because an arbitrary point $Q_{Nr}$ on the line segment $Q_{N1}Q_{N2}$ is a linear function of the stochastic vectors $Q_{N1}$ and $Q_{N2}$ and it is further assumed that $Q_{N1}$ and $Q_{N2}$ are independent, each with the same variance-covariance matrix and distributed according to an N-dimensional normal distribution, thus $Q_{Nr}$ follows the N-dimensional normal distribution:

$$Q_{Nr} \sim N_N(\mu_{1r}, \sigma_1^2, \ldots, \sigma_N^2), ((1 - r)^2 + r^2)$$

3.2.2. The confidence space of an N-dimensional line segment

To describe error of a line segment in N-dimensional space, we derive a space around the N-dimensional measured line segment $Q_{N1}Q_{N2}$. The confidence space $J_N$ of an N-dimensional line segment is such that all points $\phi_{Nr}$ with $r \in [0, 1]$ are contained within $J_N$ with a probability that is larger than a prescribed confidence level $\gamma$, (equation 10).

$$P(\phi_{Nr} \in J_{Nr} \text{ for all } r \in [0, 1]) > \gamma$$

$J_N$ is the union of sets $J_{Nr}$ for all $r \in [0, 1]$. One $J_{Nr}$ is a set of points $(x_1, x_2, \ldots, x_i,$
..., $x_N$)\textsuperscript{T} with $x_i$ ($i=1, ..., N$) satisfying

$$X_{ir} - c_{Ni} \leq x_i \leq X_{ir} + c_{Ni} \quad (11)$$

where

$$c_{Ni} = [k((1-r)^2 + r^2)]^{1/2} \sigma_i \quad (i=1, ..., N) \quad (12)$$

The parameter $k$ is dependent on the selected confidence level $\gamma$ and can be looked up in a chi-square distribution table, $k = \chi^2_{2(N-1)+\gamma}/N$. This is the case for an N-dimensional line segment.

With a risk expressed by the confidence level, we can state that the true location of the line segment is somewhere inside the confidence space. Consequently, when doubting the positional certainty of an N-dimensional feature, we are not concerned with regions outside the confidence space but only about that contained within the space as defined by equations (10) to (12). This can be utilized in analyzing error of an N-dimensional line segment. The size of the confidence space is dependent on the error of the endpoints and the predefined confidence level.

3.3. The confidence space of an N-dimensional line

The confidence space of an N-dimensional line is defined as a sub-space that includes the true location of the line with the probability larger than a predefined confidence level. Let an N-dimensional line $(Q_1Q_M)$ be composed of a sequence of N-dimensional line segments: $Q_0Q_1$, $Q_1Q_2$, ..., $Q_{M-1}Q_M$. Based on the above result of the confidence space of a line segment, we can obtain confidence spaces for each of the line segments and they are indicated as $S_1$, $S_2$, ..., $S_M$ respectively. According to the definition of the confidence space of a line and with the condition that error of the points between the line segments is independent, the confidence space of the line $Q_1Q_M$ is the union of confidence spaces of the line segments (equation 13).

$$S_{1M} = \bigcup_{j=1}^{M} (S_{(j)})(j=1, ..., M) \quad (13)$$

Let the confidence spaces of all line segments be obtained using the same confidence level, e.g. 95%. The confidence level of the confidence space for the line $Q_1Q_M$ is the same as that for all line segments, i.e., 95%.

Next, we will develop error models for one- and two-dimensional cases based on the generic models for N-dimensional case. The error indicators, corresponding to the confidence space for N dimensions, are defined as the confidence interval and confidence region for the one- and two-dimensional features in GIS respectively.

4. The positional error of one-dimensional features

4.1. A one-dimensional point

In one-dimensional space, a point is defined as a measured one-dimensional stochastic vector $Q_{10} = [X_{10}]$ which follows a one-dimensional normal distribution ($N_1$):

$$Q_{10} = [X_{10}] \sim N_1 [\mu_{10}, \sigma_{10}^2] \quad (14)$$

where $\mu_{10}$ and $\sigma_{10}^2$ are the parameters that define the statistical properties of the point.
4.1.1. The confidence interval of a point

The confidence interval \( J_{10} \) of the point is such that a random interval around the measured point \( \phi_{10} \) contains the true location of the point with a probability that is larger than a prescribed confidence level \( \gamma \) (equation 15).

\[
P(\phi_{10} \in J_{10}) > \gamma \tag{15}
\]

\( J_{10} \) is a set of points \([x_1]\) which satisfy

\[
X_{10} - a_{11} \leq x_1 \leq X_{10} + a_{11} \quad \text{(16)}
\]

where

\[
a_{11} = k_{1/2} \sigma_1 \tag{17}
\]

\( k = \chi^2_{1: \gamma} \). The parameter \( k \) is dependent on the selected confidence level \( \gamma \) that can be found in a chi-square distribution table. For example, if the confidence level \( \gamma \) is equal to 0.97, the chi-square parameter \( k \) is equal to 4.789. This is shown in figure 2a.

4.2. A one-dimensional line segment

In one-dimensional space, two points define a line segment. They are the two measured one-dimensional stochastic vectors \( Q_{11} = [X_{11}] \) and \( Q_{12} = [X_{12}] \) which follow the one-dimensional normal distributions \((N_1)\):

\[
Q_{11} = [X_{11}] \sim N_1 [\mu_{11}, \sigma^2_1] \quad \text{(18)}
\]

\[
Q_{12} = [X_{12}] \sim N_1 [\mu_{12}, \sigma^2_1] \quad \text{(19)}
\]

where \( \mu_{11}, \mu_{12} \) and \( \sigma^2_1 \) are parameters which define the statistical properties of the two endpoints.

An arbitrary point on the true location of the line segment \( \phi_{11} \phi_{12} \) is represented by

\[
\phi_{1r} = (1 - r)\phi_{11} + r\phi_{12} = [(1 - r)\mu_{11} + r\mu_{12}] \quad r \in [0, 1] \tag{20}
\]

4.2.1. The confidence interval of a one-dimensional line segment

The confidence interval \( J_1 \) of the one-dimensional line segment \( Q_{11} Q_{12} \) is a random interval such that all points \( \phi_{1r} \) with \( r \in [0, 1] \) are contained within \( J_{1r} \) with a probability that is larger than a prescribed confidence level \( \gamma \) (equation 21).

\[
P(\phi_{1r} \in J_{1r} \text{ for all } r \in [0, 1]) > \gamma \tag{21}
\]

\( J_1 \) is the union of sets \( J_{1r} \) for all \( r \in [0, 1] \). One \( J_{1r} \) is a set of points \( x_1 \) that satisfy

\[
X_{1r} - c_{11} \leq x_1 \leq X_{1r} + c_{11} \quad \text{(22)}
\]

where

\[
c_{11} = [k((1 - r)^2 + r^2)]^{1/2} \sigma_1 \quad \text{(23)}
\]

The parameter \( k = \chi^2_{1: \gamma} \) is dependent on the selected confidence level \( \gamma \) and can be looked up in a chi-square distribution table. For instance, if \( \gamma = 0.97 \), \( k = 7.174 \). The confidence interval of a one-dimensional line segment is illustrated in figure 2b.

4.3. The confidence interval of a one-dimensional line

In one-dimensional space, let a line \((Q_1 Q_M)\) be a composite of a sequence of line segments: \( Q_0 Q_1, Q_1 Q_2, ..., Q_{M-1} Q_M \). The confidence intervals for each of the line
Figure 2. (a) The confidence interval of a one-dimensional point. The confidence interval of the measured point $Q_{10}$ is indicated by the interval $J_{10}$. The true location of the measured point $Q_{10}$ is $\phi_{10}$. The confidence interval $J_{10}$ contains the true location $\phi_{10}$ with probability larger than the 97% confidence level. (b) The confidence interval of a one-dimensional line segment. The confidence interval of the measured line segment $Q_{11}Q_{12}$ is indicated by the interval $J_{1}$. The true location of the measured line segment $Q_{11}Q_{12}$ is $\phi_{11}\phi_{12}$. The confidence interval $J_{1}$ contains the true location $\phi_{11}\phi_{12}$ with a probability that is larger than the 97% confidence level. (c) The confidence interval of a one-dimensional line. The measured line $Q_{0}Q_{m}$ is composed of the line segments $Q_{0}Q_{1}, Q_{1}Q_{2}, \ldots, Q_{m-1}Q_{m}$. $I_{m}$ is the confidence interval of the line $Q_{0}Q_{m}$. It is constructed by the union of the confidence intervals of the composite line segments, i.e., $I_{1}, I_{2}, I_{3}, \ldots, I_{m}$. The true location of the measured line $Q_{0}Q_{m}$ is contained within $I_{m}$ with a probability that is larger than the 97% confidence level.

segments are $I_{1}, I_{2}, \ldots, I_{M}$ respectively. The confidence interval for the line $Q_{1}Q_{M}$ is the union of all confidence intervals of the composite line segments, i.e.

$$I_{1M} = \bigcup_{j=1}^{M} (I_{(j)}) \quad (j=1, \ldots, M)$$

This result is illustrated in figure 2c. Let all confidence intervals of the composite line segments be obtained using the same confidence level, e.g. 97%. The confidence level of the confidence interval of the line $Q_{1}Q_{M}$ is the same as for all the line segments, i.e. 97% confidence level for the line.

For a one-dimensional point, its confidence interval can be applied to describe its positional error. The confidence interval of a one-dimensional line can be used to describe, for example, the error of a part of a road section that is determined by two road section points in a one-dimensional space of a GIS. The confidence interval
of a line can be applied to describe the error that is a composite of several line segments of a road section. These descriptions give the intervals within which the true locations of the road section, points, line segments and lines, are included with a probability that is larger than a predefined confidence level.

5. Positional error of two-dimensional features

5.1. A two-dimensional point

In two-dimensional space, a point is described as a measured two-dimensional stochastic vector \( Q_{20} = [X_{10}, X_{20}]^T \) which follows a two-dimensional normal distribution \( (N_2) \):

\[
Q_{20} = \begin{bmatrix} X_{10} \\ X_{20} \end{bmatrix} \sim N_2 \left[ \begin{bmatrix} \mu_{10} \\ \mu_{20} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right] \tag{25}
\]

where \( \mu_{10}, \mu_{20}, \sigma_1^2, \sigma_2^2, \) and \( \sigma_{12} = \sigma_{21} \) are the parameters which determine the statistical properties of the measured point \( Q_{20} \).

5.1.1. The confidence region of a two-dimensional point

The confidence region \( J_{20} \) of the measured point \( Q_{20} \) is defined as a region such that the true location of the point \( (\phi_{20}) \) is contained within the region \( J_{20} \) with a probability that is larger than a prescribed confidence level \( \gamma \) (equation 26).

\[
P(\phi_{20} \in J_{20}) > \gamma \tag{26}
\]

The confidence region \( J_{20} \) is determined by a set of points \( (x_1, x_2)^T \) with \( x_1 \) satisfying

\[
X_{10} - a_{21} \leq x_1 \leq X_{10} + a_{21} \tag{27}
\]

and \( x_2 \) satisfying

\[
X_{20} - a_{22} \leq x_2 \leq X_{20} + a_{22} \tag{28}
\]

where

\[
a_{21} = k^{1/2} \sigma_1 \tag{29}
\]

and

\[
a_{22} = k^{1/2} \sigma_2 \tag{30}
\]

The parameter \( k \) is dependent on the selected confidence level \( \gamma \) and can be looked up in a chi-square distribution table, \( k = \chi^2_{1;1+\gamma/2} \). For example, if \( \gamma = 0.97 \), \( 1 + \gamma/2 = 0.985 \), \( k = 6.024 \). This result is illustrated in figure 3a.

5.2. A two-dimensional line segment

In two-dimensional space, we denote two points that define a two-dimensional line segment by the two measured stochastic vectors \( Q_{21} = [X_{11}, X_{21}]^T \) and \( Q_{22} = [X_{12}, X_{22}]^T \). The two stochastic vectors follow the two-dimensional normal distributions \( (N_2) \):

\[
Q_{21} = \begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix} \sim N_2 \left[ \begin{bmatrix} \mu_{11} \\ \mu_{21} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right] \tag{31}
\]

\[
Q_{22} = \begin{bmatrix} X_{12} \\ X_{22} \end{bmatrix} \sim N_2 \left[ \begin{bmatrix} \mu_{12} \\ \mu_{22} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right] \tag{32}
\]
Figure 3. (a) The confidence region of a two-dimensional point. $Q_{20}$ is a measured two-dimensional point. $\phi_{20}$ shows the true location of the measured point $Q_{20}$. The confidence region of the measured point $Q_{20}$ is a rectangle with sides $2Q_{21}$ and $2Q_{22}$. The true location of the point ($\phi_{20}$) is contained within this confidence region with a probability that is larger than the 97% confidence level. (b) The confidence region of a two-dimensional line segment. $Q_{21}Q_{22}$ is a measured two-dimensional line segment. $\phi_{21}\phi_{22}$ indicates the true location of the measured line segment $Q_{21}Q_{22}$. The confidence region of the measured line segment is the region indicated by $J_2$. The true location of the line segment ($Q_{21}Q_{22}$) is contained within the confidence region with a probability that is larger than the 97% confidence level. (c) The confidence region of a two-dimensional line. The confidence region of the line is the shaded area which is constructed by the union of the confidence regions of the composite line segments. The regions are indicated by $R_1$, $R_2$, $R_3$, $R_M$. The true location of the line is contained within the region with a probability that is larger than the 97% confidence level.
where $\mu_{11}$, $\mu_{21}$, $\mu_{12}$, $\mu_{22}$, $\sigma_1^2$, and $\sigma_2^2$, $\sigma_{12} = \sigma_{21}$, are the parameters which define the statistical properties of the two endpoints.

5.2.1. The confidence region of a two-dimensional line segment

The confidence region $J_2$ of a two-dimensional line segment is such that all points $\phi(r)$ with $r \in [0, 1]$ are contained within $J_2$ with a probability that is larger than a prescribed confidence level $\gamma$ (equation 33).

$$P(\phi(r) \in J_2 \text{ for all } r \in [0, 1]) > \gamma$$

($J_2$ is the union of sets $J_{2r}$ for all $r \in [0, 1]$). One $J_{2r}$ is a set of points $(x_1, x_2)^T$ with $x_1$ satisfying

$$X_{1r} - c_{21} \leq x_1 \leq X_{1r} + c_{21}$$

and $x_2$ satisfying

$$X_{2r} - c_{22} \leq x_2 \leq X_{2r} + c_{22}$$

where

$$c_{21} = [k((1 - r)^2 + r^2)]^{1/2} \sigma_1$$

and

$$c_{22} = [k((1 - r)^2 + r^2)]^{1/2} \sigma_2$$

The parameter $k$ is dependent on the selected confidence level $\gamma$, such that $k = x_{2,11 + \gamma/2}$ (Shi 1994). The value of $\chi^2$ from a chi-square distribution table. For example, if $\gamma = 0.97$, $(1 + \gamma)/2 = 0.985$, $k = 8.517$ (Mikhail and Ackermann 1976).

5.2.2. The shape of the confidence region of a two-dimensional line segment

To analyse the shape of the confidence region of a two-dimensional line segment, we differentiate the equations (36) and (37). With the conditions $\partial c_{21}/\partial r = 0$ and $\partial c_{22}/\partial r = 0$, $c_{21}$ and $c_{22}$ get their minimum value at $r = 1/2$. The other two extreme points of $c_{21}$ and $c_{22}$ are at $r = 0$ and 1. This means that the confidence region is narrowest at the centre point of the line segment ($r = 1/2$), and widest at the two endpoints of the line segment ($r = 0$ or 1). The shape of the confidence region of a two-dimensional line segment is shown in figure 3b.

Based on the analytical form of the confidence region of a line segment, it can be stated that the size of the confidence region is dependent on the error of endpoints and the predefined confidence level. The larger the error of endpoints and the predefined confidence level, the larger the confidence region. The shape of the confidence region is smaller at the middle part and larger around the two endpoints.

5.3. The confidence region of a two-dimensional line

In two-dimensional space, let a two-dimensional line ($Q_1Q_2...Q_M$) be composed of a sequence of two-dimensional line segments: $Q_0Q_1$, $Q_1Q_2$, ..., $Q_{M-1}Q_M$. The confidence regions of each of the line segments are described as $R_1$, $R_2$, ..., $R_M$ respectively. The confidence region of a line is defined as a region around the measured line within which the true location of the line is included with a probability larger than a predefined confidence level. With the condition that the error of the points between
the line segments is independent, the confidence region of the line \( Q_1 Q_M \) is the union of all confidence regions of the composite line segments (equation 38).

\[
R_{1M} = \bigcup_{j=1, \ldots, M} (R_{ij})
\]

(38)

Figure 3c illustrates the confidence region of the line \( Q_1 Q_M \). Let the confidence regions of all line segments be obtained using the same confidence level, e.g., 97%. The confidence level of the confidence region of the line \( Q_1 Q_M \) is the same as that for all line segments, i.e., 97%.

6. Conclusions and discussions

In this paper, a newly-developed fundamental approach for modelling error of GIS data has been presented. The approach is to model positional error of geometric features in GIS, from one- to N-dimensional space. The geometric features are points, line segments and lines. With such an approach, one can identify a certain space within which the true location of a geometric feature is included with a probability larger than a predefined confidence level. The size and shape of the space are determined by the error of the composite points and the predefined confidence level. These spaces are defined as the confidence interval, confidence region and confidence space for the one-, two- and N-dimensional geometric features in GIS respectively.

Compared with existing error models for features in GIS, the approach presented in this paper gives a generic solution for N-dimensional space. These models can be directly applied to model the error of geometric features in the existing one-, two- and three-dimensional GIS and can be expanded to N-dimensional models as necessary, e.g. spatio-temporal GIS. Under the assumptions that errors of composite points are independent and follow multi-dimensional normal distributions, the models presented in this paper are derived using rigorous statistical methods. This has the advantage of being easily implemented in the existing GIS, because the models are formulated mathematically.

An important application of this model is the analysis of error propagation in the GIS based spatial analysis, such as an overlay. First, the confidence region models for point, line segment and lines are applied to the GIS features in each of the layers. Second, the percentages of the total area covered by all confidence regions of GIS features in each layer are calculated. This is a positional error indicator for the GIS features in that layer. Third, the confidence regions in each layer can be further combined by an overlay operation in the GIS. The percentage of the area covered by all confidence regions in the combined layer is further calculated and used as a positional error indicator for the final combined layer. From the comparison of the total area of the confidence region in the final combined layer with that of the original layers, the error propagation is derived.

Acknowledgment

This research has been funded by the Competitive Earmarked Research Grant (No. HKP 65/95E), Research Grant Council, Hong Kong, PR China.

References


