An-Najah Univ. J. Res., Vol. 11, (1997) 23-38

Modeling Positional Uncertainty of Linear Features in Geographic Information Systems

Fayez Shahin

Water & Environmental Studies Center, An-Najah N. Univ., Nablus, Palestine

Abstract

This paper describes a probabilistic approach to model positional uncertainty of linear features in a vector-based geographic information system (GIS). Positional uncertainty is one of the components of uncertainty inherent in any object description in GIS. With a number of assumptions, the positional error of an arbitrary point on a line segment is derived based on the distribution of errors at the end points of the segment. This defines the probability density and the confidence region of a line segment and a set of indicators for the error of a line segment. The union of the confidence regions of the line segments establishes the confidence region of a linear feature. The derived uncertainty model is computationally feasible and has a great promise for efficient implementation in several GIS applications.

1. Introduction

The collection and storage of vast amounts of spatially related data in geographic information systems (GIS) has far outstripped our ability to assess their reliability. For example, large areas of Earth's surface are mapped with little chance of substantial ground truth. Once entered into a GIS, calculations such as map overlays or intersections are carried out assuming that the data are accurate when it is known that this is not the case. Thus, there is a need for methods for assessing, representing, and transmitting uncertainty through calculations with maps and GIS layers so that decision makers have some idea of the reliability of their information.

In a GIS, objects (features) are generally defined by two basic types of information:(1) Spatial information describing the location and shape of geographic features and their spatial relationships to other features; and (2) descriptive (attribute) information about the features. Locational (positional) information of objects in a GIS is represented by points for features such as wells and telephone poles; lines (arcs) for features such as roads, streams and pipelines; and areas or polygons for features such as lakes, parcels and city boundaries.

Both positional and attribute values of an object carry errors due to errors in surveying, scanning and digitization. In order to measure uncertainty in GIS, it is necessary to have an error model which combines positional and attribute errors (Goodchild et al., 1992). Modeling positional uncertainty is a key issue in developing a reliable GIS, as it would open the door for better understanding of positional issues in many practical applications such as feature extraction from digital images in softcopy stations, utilities information systems and automated mapping (AM/FM). This paper discusses in detail the uncertainty model of the positional error component and its potential applications in a vector-based GIS.

Points and lines are the geometric primitives in a vector-based GIS. Research on the accuracy of points has a very long history in surveying and mapping, whereas it was only because of GIS that research was initiated on the positional uncertainty of a line. The concept of an epsilon-band enveloping a line can be traced to Perkal (Chrisman, 1982). The epsilon band has been defined as a constant distance from either side of the line and from its two end points. This model is used under the assumption that the errors are systematic and could be accurately determined. Honeycutt (Lunetta et al., 1992) discussed the epsilon-band based on cartographic generalization and probabilities of ground location within and near the epsilon-band. Dutton (1992) simulated the distribution of line segments. Caspary and Scheuring (1992) discussed the error-band, deriving the random error of a point on a line segment. Burrough and Heuvelink (1992) discussed the use of Boolean and fuzzy logical modeling to uncertainty data.

Existing work on error propagation can be found in Haining and Arbia (1993), and Wesseling and Heuvelink (1993). This existing work deals primarily with the propagation of quantitative attribute errors through

calculations. Kiiveri (1997) developed models which allow the user to make probability statements about the presence of various attributes (map classes) at given points on the ground.

However, most of the previous work does not deal with positional uncertainty and its associated calculations of accuracy standards and error indicators. Further development is needed to be able to (1) define a boundary region of an area feature where there is uncertainty whether a point belongs to this feature or to the adjacent one, (2) assign probabilities of a point within the uncertainty zone belonging to either of the two adjacent objects, and (3) define a set of error indicators for line segments that can be used for specifications and quality parameters in a GIS. Under the assumption of normally distributed errors of coordinates, the followings can be derived: (1) the confidence region of a line (2) the probability distribution of a point on the line segment in the direction perpendicular to the line segment which can be used to compute probability vectors, and (3) segment error indicators analogous to the linear and circular errors used in mapping.

2. Positional Uncertainities of Points in GIS

In a vector-based GIS, objects are categorized into point, line, and area features. An area feature is geometrically described by its boundary polygon, thus a closed linear feature. Any linear feature is composed of one or more line segments; a line segment being a straight line connecting two points (vertices, nodes). A point is geometrically described by coordinates, p(x,y), in a 2D-GIS. Consequently, errors in coordinates constitute one of the components of positional uncertainty in a GIS. The second component is caused by sampling and approximation of a curved line feature by a sequence of straight line segments. This error is directly associated with the line segment and will not be considered in this paper.

2.1 Errors of a Point

The coordinates of points in GIS are usually the result of measurements and various processing steps. Each operation involved adds an error. These errors can be classified into three groups: blunders, systematic and random errors. This paper assumes that the coordinates of a point are free of systematic errors, thus, it only deals with random errors. If the final coordinates of a point are

expressed as a function of the original measurements, the error characteristics of any point can be determined analytically using error propagation. Furthermore, the coordinate errors are assumed to follow a normal distribution and the errors of any two points are uncorrelated.

A line segment can be defined by two points $Z_1 = (X_1, Y_1)$ and $Z_2 = (X_2, Y_2)$. The two points represent a stochastic vector, following the normal distribution:

$$Z_1 = \begin{bmatrix} X_1 \\ Y_1 \end{bmatrix} \approx N_2\left(\begin{bmatrix} \mu_1 \\ \nu_1 \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix}\right)$$

$$Z_2 = \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} \approx N_2\left(\begin{bmatrix} \mu_2 \\ \nu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix}\right) \qquad (1)$$

The expectations (true values) of points Z_1 and Z_2 are $\zeta_1 = (\mu_1, \nu_1)$ and $\zeta_2 = (\mu_2, \nu_2)$ respectively. The variance of a point in the X direction is σ_{xx} and that in the Y direction is σ_{yy}. The covariance in the XY direction is σ_{xy}. It should be noted that equal (co-)variances at the end points of a segment are assumed, which may not be realistic in the case of a multi-source GIS.

3. *Positional Uncertainity of a Line Segment*

A line segment (Figure 1) is defined by an arbitrary point (ζ_r) of the straight line connection of the end points ζ_1 and ζ_2 as

$$\zeta_r = (1 - r)\,\zeta_1 + r\,\zeta_2 \qquad \text{for } 0 \le r \le 1 \qquad (2)$$

Equation 2 shows that Z_r (the realization of ζ_r) is a linear function of Z_1 and Z_2, thus, Z_r is also normally distributed:

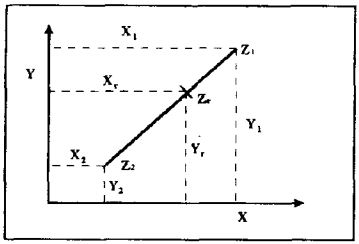

Figure 1: *Definition of a Line Segment*

$$Z_r = \begin{bmatrix} X_r \\ Y_r \end{bmatrix} \approx N_2\left(\begin{bmatrix} (1-r)\mu_1 + r\mu_2 \\ (1-r)\nu_1 + r\nu_2 \end{bmatrix} , ((1-r)^2 + r^2) \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix} \right) \qquad (3)$$

3.1 Perpendicular Distribution

The probability density function of a point in an arbitrary direction can be derived as the marginal probability density of the bivariate distribution of the point. To obtain the probability distribution in the direction perpendicular to the line segment, the original XY coordinate system is rotated to the X′Y′ by an angle θ such that the X′ axis is parallel to the line segment $\zeta_1\zeta_2$. Let Z_r' be the transformed random vector of the line Z_r, Z_r' is again normally distributed because the rotation of the axis is a linear transformation. The marginal probability density of Z_r' in the Y′ direction results as

$$f_{y'}(y') = \int_{-\infty}^{\infty} f(x', y')dx' = \frac{1}{(2\pi)^{0.5}(\sigma_{yy'})^{0.5}} \exp(-(y' - \nu_r')^2 / 2\sigma_{yy'}) \qquad (4)$$

where: $E(Y_r') = \nu_r' = -\sin(\theta)((1-r)\mu_1 + r\mu_2) + \cos(\theta)((1-r)\nu_1 + r\nu_2)$

$$\sigma_{yy'} = [A(-\sin(\theta)) + B(\cos(\theta))]((1-r)^2 + r^2)$$

$A = \cos(\theta)\,\sigma_{xy} - \sin(\theta)\sigma_{xx}$

$B = \cos(\theta)\sigma_{yy} - \sin(\theta)\sigma_{xy}$

θ is the rotation angle from the XY system to the X'Y' system

3.2 Probability Distribution of a Line Segment

The distribution of the line segment Z_1Z_2 can be characterized by three density functions: the perpendicular density for any $r \in (0,1)$ and the density at the two end points Z_1' and Z_2'. The probability density function of Z_t' (t=1 or 2) is:

$$f_t'(x',y') = \frac{1}{\sqrt{2\pi}\left|\sum'_t\right|^{0.5}} \exp(-0.5(Z_t' - E(Z_t'))^T(\sum'_t)^{-1}(Z_t' - E(Z_t'))) \qquad (5)$$

where: $Z_t' = \begin{bmatrix} X_t' \\ Y_t' \end{bmatrix}$

$$R = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

$$\Sigma_t' = \begin{bmatrix} \sigma'_{xx} & \sigma'_{xy} \\ \sigma'_{xy} & \sigma'_{yy} \end{bmatrix} = R \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix} R^T$$

$$E(Z_t') = \begin{bmatrix} \mu_t' \\ \nu_t' \end{bmatrix} = R \begin{bmatrix} \mu_t \\ \nu_t \end{bmatrix}$$

Common GIS operations that involve calculation of lengths, perimeters and areas of features need assessment of the accuracy of their results. Given the probability distribution of an arbitrary point on a line segment, means and variances of the results of such operations can be computed. For example,

this model can be used to compute the total area and its variance of a particular soil type, vegetation cover or property holding.

3.3 Confidence Region of a Line Segment

The confidence region of a line segment Z_1Z_2 is a region around the segment such that it covers the true location of this line with a predefined probability. The derivation of the confidence region is based on the distribution of an arbitrary point on the line segment. If the variance matrix of Z_r is known, a chi-square distributed statistics of X_r and Y_r and then a confidence region J_r for ζ_r can be derived. The confidence region J_r can be constructed such that it contains ζ_r with a predefined confidence level γ, while also all other ζ of the line segment are contained in their respective confidence regions. This involves an upper bound condition, leading to the inequality

$$P(\zeta_r \in J_r,\ r \in [0,1]) > \gamma \qquad (6)$$

The confidence region J of a line segment is the union of sets J_r for all $r \in [0,1]$. One region J_r is a set of points $(x,y)^T$ satisfying

$$X_r - c \le x \le X_r + c \qquad (7)$$

$$Y_r - d \le y \le Y_r + d \qquad (8)$$

where:

$$c = k^{0.5}\,[((1-r)^2 + r^2)\,\sigma_{xx}]^{0.5}$$

$$d = k^{0.5}\,[((1-r)^2 + r^2)\,\sigma_{yy}]^{0.5}$$

The parameter k is dependent on the selected confidence level γ and can be obtained from a chi-square table.

According to equations 7 and 8, the values of the parameters c and d are largest at the end points of a line segment ($r = 0$ or 1) and smallest at the center point ($r = 0.5$). Therefore, the confidence region is smallest at the center of the line segment and largest at the end points. Figure 2 shows the shape of the confidence region of a line segment.

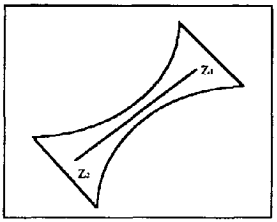

Figure 2: *Confidence Region of a Line Segment.*

The confidence region of an area feature can be utilized in the analysis of polygon overlays (Figure 3). Polygon over lays is a spatial operation which overlays one polygon layer onto another to create a new polygon layer. By creating confidence regions for the overlaid polygon layers, confidence regions for the new layer are derived. The derived confidence regions give an indication for the accuracy of the new obtained layer. Another direct application is determining an overall measure of positional uncertainty of a line map or a vector-based GIS. Around all line segments in the GIS or digital map, a confidence region can be created to compare the sum of the areas covered by the confidence regions with the total area of this map. The ratio of the areas of the map covered by the confidence regions to the total area of the map is an error indicator for the positional uncertainty of the map with line features.

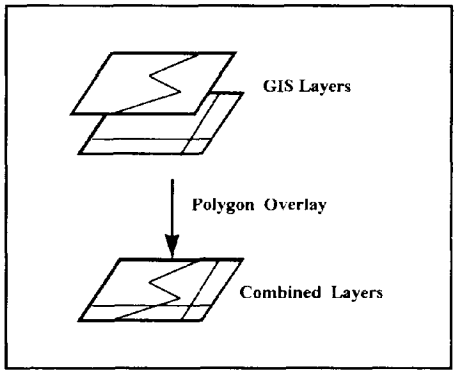

Figure 3: *Area Features in Polygon Overlays.*

4. Segment Error Indicators

In surveying and mapping, a number of uncertainty indicators were defined for one-dimensional random variables (linear errors) and for two-dimensional variables (circular errors) (see Canadian Council, 1982). It may be useful to extend these concepts to a line segment. Based on the probability density of a line segment (equation 5), a set of indicators for segment errors can be defined (Figure 4). These segment errors include the segment standard error, segment probable error, segment map accuracy standard and segment near certainty error.

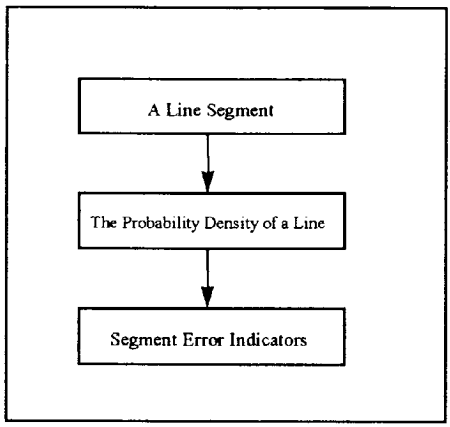

Figure 4: *Segment Error Indicators.*

4.1 Segment Standard Error

The quantity $\sigma_b(r)$ for all $r \in [0,1]$ is called the segment standard error. It can serve as the basic statistical parameter indicating the spread of a line segment. $\sigma_b(r)$ is defined as:

$$\sigma_b(r) = (\sigma((1-r)^2 + r^2))^{0.5} \qquad (9)$$

where

$$\sigma = (\cos(\theta)\sigma_{xy} - \sin(\theta)\sigma_{xx})(-\sin(\theta)) + (\cos(\theta)\sigma_{yy} - \sin(\theta)\sigma_{xy})(\cos(\theta))$$

When the statistical parameters of the two end points and the rotation angle θ of the line segment are fixed, σ is a constant value. $\sigma_b(r)$ takes the maximum value $(\sigma)^{0.5}$ at the end points ($r = 0$ or 1) and its minimum $(\sigma/2)^{0.5}$ at midpoint ($r = 0.5$). A region is constructed by drawing a curve on each side of $\zeta_1\zeta_2$ at distances of $\sigma_b(r)$, r ranging from 0 to 1. According to the table of the one-dimensional normal distribution, this region occupies an area

of 68% of the total area covered by the density surface of the line segment. An alternative name for the segment standard error is root mean square error of a segment.

4.2 Segment Probable Error

A segment probable error is defined as the fraction of the segment standard error that delineates a distribution region being equal to 50% of the total area covered by the density surface. According to the table of the one-dimensional normal distribution, the segment probable error SPE(r) is:

$$SPE(r) = 0.675\ \sigma_b(r) \quad \text{for all } r \in [0,1] \qquad (10)$$

4.3 Segment Map Accuracy Standard

The segment map accuracy standard is defined by the distribution region, being 90% of the total area covered by the density surface. According to the normal distribution table, the segment map accuracy standard SMAS(r) is:

$$SMAS(r) = 1.645\ \sigma_b(r) \quad \text{for all } r \in [0,1] \qquad (11)$$

4.4 Segment Near Certainty Error

The segment near certainty error is defined by a distribution region of 99.74% of the total area covered by the density surface. According to the normal distribution table, the segment near certainty error SNCE(r) is:

$$SNCE(r) = 3.0\ \sigma_b(r) \quad \text{for all } r \in [0,1] \qquad (12)$$

The new derived segment error indicators are listed in table 1. They represent an extension of the linear and circular error indicators. The segment standard error could be further extended to also include the sampling error mentioned before. The segment standard error or any of its derived quantities offer a manageable descriptor of the accuracy of line segments. In a GIS, it could be included as a quality parameter in either the meta data in a homogeneous data base, or attached to every line segment in a multi-source data.

Term	Symbol	P%
Segment Standard Error	$\sigma_b(r)$	68.26
Segment Probable Error	SPE(r)	50
Segment Map Accuracy Standard	SMAS(r)	90
Segment Near Certainty Error	SNCE(r)	99.74

Table 1: Segment Error Indicators

5. *Positional Uncertainty of Linear Features*

A linear feature is very often composed of several line segments. By adequately combining the uncertainty of the line segments, the confidence region of an entire feature can be easily obtained as well as unique probability values at the joints of segments. The confidence region of a linear feature, whether open or closed, can be generated as the union of the confidence regions of the constituent line segments. It provides an uncertainty zone around an object (Figure 5).

The problem of how to determine the probability that a point in the region of influence of joining line segments belongs to an object can be solved using fuzzy subset theory. Let the probability that Q belongs to object S based on the distribution of line segment L_1 be $P_1(Q \in A)$; the probability that Q belongs to object S based on the distribution of line segment L_2 be $P_2(Q \in A)$. According to the fuzzy subset theory, the probability that point Q belongs to object S, P ($Q \in A$), is :

$$P(Q \in A) = min\,[P_1(Q \in A), P_2(Q \in A)] \qquad (13)$$

This means that the minimum operation between segments L1 and L2 in the joint region can be utilized to generate unique uncertainty values for composed linear features. Therefore, this minimum operation could be implemented in a practical GIS to determine the category (label) of a point on the boundaries of two or more classes (parcels, soil types or vegetation

cover). Determining the label of a point on the boundaries of two or more classes is a very common problem in GIS. Accordingly, this probabilistic approach has a potential for efficient integration in a GIS.

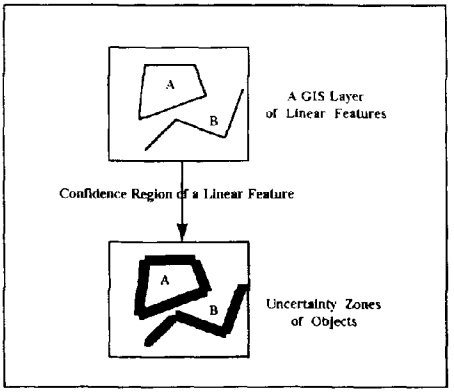

Figure 5: *Uncertainty Zones of Objects in a GIS.*

6. Conclusion

The paper discussed in detail positional uncertainty of linear features in a vector-based GIS. The uncertainty of a line segment has been established as a key issue. With a number of simplifying assumptions, an analytical derivation of an uncertainty zone and uncertainty values has been demonstrated. Based on the derived confidence region and probability distribution of line segments, the uncertainty of a linear feature in a GIS was described and a set of segment error indicators was defined as an extension of linear and circular error indicators. The derived models and error indicators are computationally feasible and have several practical applications. Therefore, this probabilistic approach has a great promise for efficient integration in a GIS.

There still are many more issues to investigate. Among them: refined approaches to utilize the probability function of a line segment for computing probability vectors, introducing tail-cut normal distribution, adding correlated

point sequences and sampling error, implementation and demonstration of typical applications.

Acknowledgment

First of all, I have to thank Prof. Munther Salah, the president of An-Najah National University, for his support and encouragement. Special thanks should also go to the director and staff of the Water and Environmental Studies Center for their cooperation. Finally, the external reviewers of this paper are appreciated for their most helpful comments.

References

Burrough, P., and Heuvelink, G. (1992). The sensitivity of Boolean and continuous (fuzzy) logical modeling to uncertainty data. In: Proceedings of EGIS'92. Munich-Germany. Vol. 1, pp. 1032-1039.

Canadian Council on Surveying and Mapping (1982). National Standards for the Exchange of Digital Topographic Data, II - Standards for the Quality Evaluation of Digital Topographic Data. Energy, Mines and Resources, Ottowa-Canada, 83p.

Caspary, W., and Scheuring, R. (1992). Error-bands as a measure of geometrical accuracy. In: Proceedings of EGIS'92. Munich-Germany. Vol. 1, pp. 226-233.

Chrisman, N. R. (1982). A theory of cartographic error and its measurement in digital data base. In: Proceedings of Auto-Carto 5. Baltimore-USA. Vol. 1, pp. 159-168.

Dutton, G. (1992). Handling positional uncertainty in spatial databases. In: Proceedings of the 5th International Symposium on Spatial data Handling. Charleston-USA. Vol. 2, pp. 460-469.

Goodchild, M.F., Sun, G., and Yang, S. (1992). Development and test of an error model for categorical data. International Journal of GIS, 6 (2), 87-104.

Haining, R. P., and Arbia, G. (1993). Error propagation through map operations. Technometrics, 35, 293-305.

Heuvelink, G. B., and Burrough, P. A. (1993). Error propagation in cartographic modeling using Boolean logic and continuous classification. International Journal of GIS, 7, 231-246.

Kiiveri, H.T. (1997). Assessing, representing, and transmitting positional uncertainty in maps. International Journal of GIS, 11(1), 33-52.

Lunetta, R.S. et al. (1992). Remote sensing and geographic information system data information: error sources and research issues. Journal of Photogrammetric Engineering and Remote Sensing, 57(6), 677-687.

نمذجة الشك (الخطأ) في موقع المعالم الخطية في نظم المعلومات الجغرافية

فايز شاهين

ملخص:

يعتبر الخطأ في موقع أي معلم من العناصر الرئيسية للأخطاء الموجودة في نظم المعلومات الجغرافية. ورقة البحث هذه تصف طريقة احتمالية لنمذجة الخطأ في موقع المعالم الخطية. على افتراض أن الخطأ في موقع نقطة على خط يعتمد على الخطأ في نقطة البداية والنهاية لهذا الخط، فإن التوزيع الاحتمالي لنقطة ما على الخط يمكن إشتقاقه من التوزيع الاحتمالي المعروف لنقطتي البداية والنهاية. بالاعتماد على التوزيع الاحتمالي المشتق للخط فإن منطقة الثقة للخط ومجموعة من المعايير يمكن أيضا اشتقاقها لهذا الخط. إن اتحاد مناطق الثقة لمجموعة من الخطوط المشكلة لمعلم خطي يشكل منطقة الثقة لهذا المعلم الخطي. هذه الطريقه الاحتمالية لها تطبيقات عمليه هامه جدا في نظم المعلومات الجغرافية.