Building the Estimation Model of Digitizing Error

Huang Youcai and Liu Wenbao

Abstract

Moving a digitizer along a smooth and continuous line could be regarded as a discrete time stochastic process consisting of trend motion and random motion. A stochastic stationary observation series of digitizing error may be generated by adopting a backward difference process (filtering the trend motion from the stochastic series). To separate the trend motion from the stochastic series efficiently, several mathematical formulae have been developed for measuring the complexity of line related to the determination of order of the backward difference operators. The stochastic motion may be simulated by using an autoregressive process in terms of time series analysis theory. The estimation model of digitizing error, consisting of these two processes, has been built. Numerical examples presented in this paper show how to use the model to estimate the digitizing error after having a set of digitized data.

Introduction

For establishing a geographic information system (GIS), a digitizing approach is one of the common processes used to create a spatial database of objects. As Goodchild and Gopal (1992) stated, positional accuracy will be affected by the operator's precision in positioning the cursor and by the rules used to select points to be digitized from line or polygon objects. Keefer *et al.* (1988) pointed out that manual digitizing is performed either in stream mode or point mode. For the stream mode, the coordinates of digitizing points are recorded at some specified, regular time, or distance interval as the cursor is moved continuously along the map line. The digitizing process in this type of mode is serially correlated; i.e., the observations are not independent. They have used autocorrelated process (stochastic time autoregressive process) to model digitizing error. Burrough (1986) has evaluated the errors of digitizing a set of discrete points (point mode) by means of comparison between their true and digitized coordinates. Unfortunately, any such model would have to be sensitive to the type of line being digitized (Goodchild and Gopal, 1992).

As Tobler (1992) described, the results of an analysis of geographical data should not depend on the spatial coordinates used — the results should be frame independent. From this point of view, this paper proposed an estimation model of digitizing error by which the result of digitizing error estimated may be invariant for different types of lines. The process of digitizing a line or a polygon could be considered as a stochastic trial (stochastic series) which contains trend and random motions in the stream mode. Removing the trend motion from the stochastic series could be an efficient way to generate the random field to which a common estimation procedure can be applied. In addition to that, the dependence of the estimation model on the type of line being digitized would be reduced substantially if the appropriate order of a polynomial for simulating this line could be found. For this purpose, several methods have been proposed for the quantitative measure of the complexity of a line, which may lead to selecting the order of the backward difference operators properly.

To build an estimation model of digitizing error, Goodchild and Dubuc (1987) and others sought to develop a procedure which can produce a "random" map. Haining *et al.* (1983) have made use of a simple spatially autoregressive process to generate a random field of known distribution and classify the results by comparing the value in each pixel to the prescribed probabilities. As far as the stream mode is concerned, the errors of digitizing a continuous line are supposed to be positively correlated between adjacent points along the digitized line. In this case, multivariate normal processes which are maximum entropy distributions for their mean vectors and covariance matrices can be used to build the error estimation model. The variance/covariance can serve as the basis for a multivariate normal approximation to the distribution (Lavenda and Scherer, 1987). The autoregressive approach might possibly, however, be used more easily to generate a string of realization themselves, due to the convenience of normal distribution generators. In mathematical modeling, it is often advantageous (conceptionally and/or computationally) to pass from a discrete framework to a continuous one.

Description of Moving Trace of Digitizing a Smooth and Continuous Line

Assume that $\{z_t = (x_t, y_t)^T, t = 1, 2, ..., n\}$ is an observation sequence consisting of coordinates of digitized points along a smooth and continuous line denoted by $f(x, y) = 0$. The $\{z_t\}$ represents the moving trace of a digitizer to be used, which may be written in the form

$$\{z_t\} = \{\bar{z}_t\} + \{w_t\} + \{\varepsilon_t\}, \tag{1}$$

where $\{\bar{z}_t = (\bar{x}_t, \bar{y}_t)^T\}$ is a function determined by the type of line $f(x,y) = 0$; $\{w_t = (w_{x_t}, w_{y_t})^T\}$ denotes the stochastic function determined by digitizer's motion, and $\{\bar{z}_t\}$ and $\{w_t\}$ stand for trend and stochastic motions along line $f(x,y) = 0$, respectively. $\{\varepsilon_t = (\varepsilon_{x_t}, \varepsilon_{y_t})^T\}$ denotes a random error sequence. In general, a topographic line $f(x,y) = 0$ can be expressed by a multiparametric function, i.e.,

$$\bar{x}_t = g(t); \qquad \bar{y}_t = h(t), \tag{2}$$

where t is a time parameter. If $f(x,y) = 0$ is smooth and continuous, it can be approximately simulated by a polynomial function. In terms of cartographic theory, parametric equations $g(t)$ and $h(t)$ can be written in an mth-order polynomial equation, i.e.,

The School of Earth Sciences and Engineering Surveying, Wuhan Technical University of Surveying and Mapping, Wuhan 430070, China.

H. Youcai is presently with the Department of Civil Engineering, University of Washington, 121 More Hall, Box 352700, Seattle, WA 98195.

Photogrammetric Engineering & Remote Sensing, Vol. 63, No. 10, October 1997, pp. 1203–1209.

0099-1112/97/6310-1203$3.00/0
© 1997 American Society for Photogrammetry and Remote Sensing

$$\bar{x}_t = \sum_{k=0}^{m} a_k t^k; \qquad \bar{y}_t = \sum_{k=0}^{m} b_k t^k, \tag{3}$$

where a_k, b_k ($k = 0,1,2,\dots, m$) are polynomial coefficients. Because manual digitizing is usually a low speed process, $\{w_t\}$ could be regarded as a smoothly stochastic two-dimensional variables with zero mean vector. The "smooth" means that the direction of a line does not change drastically and also that the data are generated from this line using a digitizer with a normal or low speed. Then two-dimensional autoregressive model may be established based on the time series analysis (Haining *et al.*, 1983).

The digitized data $\{z_t\}$ could be considered as a stochastic observation sequence with equal accuracy if the condition of digitizing remains unchanged. The mathematical expectation of the observations may be different because the digitizer's positions along a line are continuously changing. From this point of view, we could consider $\{z_t\}$ as a normally distributed random variables vector with mean vector μ_t and covariance matrix Γ (Caspary and Scheuring, 1993), i.e., $\{z_t\} \sim N(\mu_t, \Gamma)$. In order to estimate the covariance matrix Γ consisting of diagonal elements $= \sigma_x^2$; σ_y^2 and off-diagonal elements $= \sigma_{xy}$, it is required to filter the low frequency part (trend motion) from the observation series by using a highpass digital filter (Herzog, 1992). As a result, a random line or curve is generated (Goodchild and Dubuc, 1987). The filtering process may be realized by employing a backward difference operator in the observation sequence (Box and Jenkins, 1976).

Definition and Characteristics of Operators

Based on the theory of the time series and system analysis, an estimation model of digitizing error can be established by adopting a set of operators (Box and Jenkins, 1976, pp. 8–16). Therefore, it is necessary to define the different types of operators to be used in this paper.

Definition of the backward difference operator: assume that $\nabla_{z_t} = z_t - z_{t-1}$ in which ∇ denotes the first-order backward difference operator and correspondingly the mth-order difference operator ∇^m satisfies $\nabla^m z_t = \nabla^{m-1}(\nabla z_t)$, particularly $\nabla^0 \equiv 1$.

Definition of the backward shift operator: assume that $Bz_t = z_{t-1}$ in which B stands for the first-order backward shift operator and correspondingly the mth-order backward shift operator satisfies $B^m z_t = B^{m-1}(Bz_t)$, particularly $B^0 \equiv 1$.

Definition of the unit matrix operator: assume that $Iz_t = z_t$ in which I denotes the unit matrix operator or the unit matrix.

Characteristic 1: ∇ is a linear operator, which satisfies linear operational regulations such as an exchangeable operation.

Characteristic 2: If function $f(x)$ is the mth-order polynomial, then $\nabla^k f(x)$ is the $(m-k)$th-order polynomial $(0 \le k \le m)$ and $\nabla^{m+k} f(x) = 0$ (k is any positive integer).

Characteristic 3: The relationship between ∇, B, and I is $\nabla = I - B$ and the mth-order backward difference operator can be expressed as

$$\nabla^m = (I - B)^m = \sum_{k=0}^{m} (-1)^{m-k} C_m^k B^{m-k}, \tag{4}$$

where $C_m^k = m!/[k!(m-k)!]$.

Estimation Model of Digitizing Error

Theoretical Basis in the Estimation Model of Digitizing Error Building

To filter the trend part from observations $\{z_t\}$, the $(m+1)$th-order backward difference operator ∇^{m+1} is applied to Equation 1 and take into account $\nabla^{m+1}\bar{z}_t = 0$, then we have

$$\{\nabla^{m+1} z_t\} = \{\nabla^{m+1} w_t\} + \{\nabla^{m+1} \varepsilon_t\}. \tag{5}$$

Reviewing Equations 1 and 5, $\{\nabla^{m+1} z_t\}$ is a stationary stochastic two-dimensional sequence with zero mean vector. Then a two-dimensional autoregressive model (Box and Jenkins, 1969, p. 9) may be written as

$$\nabla^{m+1} z_t - \phi_1 \nabla^{m+1} z_{t-1} - \cdots - \phi_p \nabla^{m+1} z_{t-p} = \mu_t, \tag{6}$$

where ϕ_i ($i = 1,2,\cdots,p$) is a 2 by 2 coefficient matrix; $\phi_0 = I_2$ and $\phi_p \neq 0$. $\{\mu_t = (\mu_{x_t} \mu_{y_t})^T\}$ is a two-dimensional white Gaussian sequence with zero mean vector; i.e.,

$$E(\mu_t) = 0; \qquad E(\mu_t \mu_s^T) = \delta_{ts} Q, \tag{7}$$

where E denotes the mathematical expectation operator; δ_{ts} is the Kronecker δ-function, $\delta_{ts} = 1$ for $t = s$, and $\delta_{ts} = 0$ for others; and Q is 2 by 2 positive definite covariance matrix. Equation 6 is called an autoregressive (AR) process of order p. Based on the theory of the time series analysis, the autoregressive operator can be defined as

$$\phi(B) = I_2 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \tag{8}$$

where I_2 is the 2 by 2 unit matrix and $\phi(B)$ is the pth-order matrix polynomial. Then Equation 6 may be written economically as

$$\phi(B)\, \nabla^{m+1} z_t = \mu_t. \tag{9}$$

If the characteristic matrix polynomial $\phi(B)$ (corresponding to the AR operator) has all its zeroes outside the unit circle (Priestley, 1981), then Equation 9 may be regarded as the stationary stochastic model. In other words, a stochastic process is said to be strictly stationary if its properties are unaffected by a change of time origin. Substituting Equation 8 for $\phi(B)$ in Equation 9 and assuming $\mu_t' = \phi(B)\nabla^{m+1} w_t$, after some arrangement, we have

$$\phi(B)\nabla^{m+1}\varepsilon_t = \mu_t - \mu_t', \tag{10}$$

which may be written in details as follows

$$\phi_0 \nabla^{m+1}\varepsilon_t - \phi_1 \nabla^{m+1}\varepsilon_{t-1} - \phi_2 \nabla^{m+1}\varepsilon_{t-2} - \cdots - \phi_p \nabla^{m+1}\varepsilon_{t-p} = \mu_t - \mu_t',$$

Applying the variance operator *Var* to both sides of Equation 10 and assuming $M = Var[\nabla^{m+1}\varepsilon_t]$, $Q' = Var[\mu_t']$, then we have

$$Var[\phi_0 \nabla^{m+1}\varepsilon_t - \phi_1 \nabla^{m+1}\varepsilon_{t-1} - \cdots - \phi_p \nabla^{m+1}\varepsilon_{t-p}] = Var[\mu_t - \mu_t']$$

and then

$$\phi_0 M \phi_0^T + \phi_1 M \phi_1^T + \cdots + \phi_p M \phi_p^T = Q + Q',$$

which may be expressed simply by

$$\sum_{i=0}^{p} \phi_i M \phi_i^T = Q + Q'. \tag{11}$$

Considering Equation 4, $\nabla^{m+1}\varepsilon_t$ may be represented by

$$\nabla^{m+1}\varepsilon_t = \sum_{k=0}^{m+1} (-1)^{m+1-k} C_{m+1}^k \varepsilon_{t+k-(m+1)}. \tag{12}$$

Applying the variance operator *Var* to both sides of Equation 12, we have

$$M = \Gamma\, [\sum_{k=0}^{m+1} (C_{m+1}^k)^2]. \tag{13}$$

Taking into account Equation 13 and assuming $\lambda = [\sum_{k=0}^{m+1} (C_{m+1}^{k})^2]$, Equation 11 may be written in the form

$$\lambda(\phi_0 \Gamma \phi_0^T + \phi_1 \Gamma \phi_1^T + \cdots + \phi_p \Gamma \phi_p^T) = Q + Q'.$$

To compute the autocovariance matrix Γ, the matrix stack operator is adopted in the above equation; then we have

$$Vec\,\Gamma = \frac{1}{\lambda}\,[\sum_{i=0}^{p} (\phi_i \otimes \phi_i)]^{-1}\,(VecQ + VecQ'), \tag{14}$$

where Vec denotes the operation which stacks one column of a matrix under the other, $\otimes$ is the Kronecker-Zehfuss product (Graferend and Sanso, 1985), and $Vec\,\Gamma = (\sigma_x^2\ \sigma_{yx}\ \sigma_{xy}\ \sigma_y^2)^T$. Because Γ, Q, and Q' are symmetrical matrices, Equation 14 can be compressed as follows:

$$Veh\Gamma = \frac{1}{\lambda}\,[\sum_{i=0}^{p} (\phi_i \underline{\otimes} \phi_i)]^{-1}\,(VehQ + VehQ'), \tag{15}$$

where $Veh\,\Gamma = (\sigma_x^2\ \sigma_{yx}\ \sigma_y^2)^T$, and Veh denotes the compressed stack operator which is employed to simplify the computation of Γ. Equation 15 is a theoretical expression of the autocovariance matrix Γ. For a stationary stochastic series, the covariance matrices defined previously are positively definite.

Estimation Model of Digitizing Error Established by Using a Set of Digitized Data

Coefficients or autoregressive parameters ϕ_i and covariances Q and Q' in Equation 15 are usually unknown. In practice, the coefficients $\phi_1, \phi_2, \cdots, \phi_p$ have to be estimated from a set of digitized data. Based on the Yule-Walker equation (Box and Jenkins, 1976, pp. 54–57) and taking into account that the theoretical autocorrelations ρ_k are replaced by the estimated autocorrelations R_k, a set of linear equations for $\phi_1, \phi_2, \cdots, \phi_p$ in terms of $R_1, R_2, \cdots, R_p$ for two-dimensional data may be obtained. That is, when both sides of Equation 6 are multiplied by $\nabla^{m+1} z_{i-k}^T$, we have

$$\nabla^{m+1} z_t\, \nabla^{m+1} z_{t-k}^T - \phi_1\, \nabla^{m+1} z_{t-1}\, \nabla^{m+1} z_{t-k}^T - \cdots - \phi_p\, \nabla^{m+1} z_{t-p}\, \nabla^{m+1} z_{t-k}^T = \mu_t\, \nabla^{m+1} z_{t-k}^T.$$

where $k = 0, \cdots, p$. Applying the mathematical expectation operator to both sides of the above equations and taking into account $E[\nabla^{m+1} z_{t-k} \nabla^{m+1} z_{t-j}^T] = R_{t-j}$ for $k \neq j$; $E[\nabla^{m+1} z_{t-k} \nabla^{m+1} z_{t-j}^T] = R_k$ for $k = j$; and $E[\mu_t\, \nabla^{m+1} z_{t-k}^T] = 0$ for $k = 0,1,2, \cdots, p$, we have

$$\begin{bmatrix} R_1^T \\ R_1^T \\ \vdots \\ R_p^T \end{bmatrix} = \begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1^T & R_0 & R_1 & \cdots & R_{p-2} \\ \vdots & \vdots & \vdots & & \vdots \\ R_{p-1}^T & R_{p-2}^T & R_{p-3}^T & \cdots & R_0 \end{bmatrix} \begin{bmatrix} \phi_1^T \\ \phi_1^T \\ \vdots \\ \phi_p^T \end{bmatrix},$$

where R_{k-j} and R_k are 2 by 2 autocorrelation matrices. Therefore, the estimated value of the parameters ϕ_i can be obtained by solving the above equation if the autocorrelations R_{k-j} and R_k are known. Covariance matrices Q and Q' may be estimated by the following equations:

$$\hat{Q} = \hat{R}_0 - \sum_{i,j=1}^{p} \hat{\phi}_j\, \hat{R}_{i-j}\, \hat{\phi}_i^T; \qquad \hat{Q}' = \hat{R}_0' - \sum_{i,j=1}^{p} \hat{\phi}_j\, \hat{R}_{i-j}'\, \hat{\phi}_i^T, \tag{16}$$

where

$$\begin{aligned} \hat{R}_k &= \frac{1}{n-m-1} \sum_{t=m+2}^{n-k} \nabla^{m+1} z_t\, \nabla^{m+1} z_{t+k}^T; \\ \hat{R}_k' &= \frac{1}{n-m-1} \sum_{t=m+2}^{n-k} \nabla^{m+1} w_t \nabla^{m+1} w_{t+k}^T. \end{aligned} \tag{17}$$

Equations 17 are formulae for computing the autocorrelations $\hat{R}_k$ and $\hat{R}_k'$ based on a set of digitized data $\{z_t\}$. Substituting $\hat{\phi}_i$ $(i = 1,2, \cdots, p)$, $\hat{Q}$, and $\hat{Q}'$ for ϕ_i, Q, and Q' in Equation 15, respectively, we have

$$Veh\hat{\Gamma} = \frac{1}{\lambda}\,[\sum_{i=0}^{p} (\hat{\phi}_i \underline{\otimes} \hat{\phi}_i)]^{-1}\,(Veh\hat{Q} + Veh\hat{Q}') \tag{18}$$

where $Veh\hat{\Gamma} = (\hat{\sigma}_x^2\ \hat{\sigma}_{yx}\ \hat{\sigma}_y^2)^T$. Equation 18 is an estimation model of digitizing error for a set of digitized data.

Simplification of the Estimation Model of Digitizing Error in the Case That Observations $\{x_t\}$ and $\{y_t\}$ Are Supposed to Be Mutually Independent

Based on the knowledge of the digitizing process along a topographic line, stochastic observation sequences $\{x_t\}$ and $\{y_t\}$ could be regarded as two independent variables; then Equation 1 can be partitioned into two equations shown as follows:

$$\begin{aligned} \{x_t\} &= \{\bar{x}_t\} + \{w_{x_t}\} + \{\varepsilon_{x_t}\}; \\ \{y_t\} &= \{\bar{y}_t\} + \{w_{y_t}\} + \{\varepsilon_{y_t}\}. \end{aligned} \tag{19}$$

Then coefficients $\hat{\phi}_i$ $(i = 1,2, \cdots, p)$ in polynomial $\hat{\phi}(B)$ become diagonal matrices $dia(\hat{\varphi}_{x_i}\ \hat{\varphi}_{y_i})$ due to $\hat{\sigma}_{yx}\ \hat{\sigma}_{xy} = 0$. In this case, Equation 18 is reduced to

$$\begin{aligned} \hat{\sigma}_x^2 &= \frac{1}{\lambda}\,(1 + \hat{\varphi}_{x_1}^2 + \hat{\varphi}_{x_2}^2 + \cdots + \hat{\varphi}_{x_p}^2)^{-1}\,(\hat{\sigma}_{\mu_x}^2 + \hat{\sigma}_{\mu_x'}^2); \\ \hat{\sigma}_y^2 &= \frac{1}{\lambda}\,(1 + \hat{\varphi}_{y_1}^2 + \hat{\varphi}_{y_2}^2 + \cdots + \hat{\varphi}_{y_p}^2)^{-1}\,(\hat{\sigma}_{\mu_y}^2 + \hat{\sigma}_{\mu_y'}^2). \end{aligned} \tag{20}$$

Equation 20 is a simplified formula for the estimation of digitizing error given that $\{x_t\}$ and $\{y_t\}$ are mutually independent. Correspondingly, Equation 16 is reduced to

$$\begin{aligned} \hat{\sigma}_{\mu_x}^2 &= \hat{r}_{x_0} - \sum_{j=1}^{p} \hat{\varphi}_{x_j}\, \hat{r}_{x_j}; \quad \hat{\sigma}_{\mu_y}^2 = \hat{r}_{y_0} - \sum_{j=1}^{p} \hat{\varphi}_{y_j}\, \hat{r}_{y_j}; \\ \hat{\sigma}_{\mu_x'}^2 &= \hat{r}_{x_0}' - \sum_{j=1}^{p} \hat{\varphi}_{x_j}\, \hat{r}_{x_j}'; \quad \hat{\sigma}_{\mu_y'}^2 = \hat{r}_{y_0}' - \sum_{j=1}^{p} \hat{\varphi}_{y_j}\, \hat{r}_{y_j}'. \end{aligned} \tag{21}$$

where

$$\begin{aligned} \hat{r}_{x_k} &= \frac{1}{n-m-1} \sum_{t=m+2}^{n-k} \nabla^{m+1} x_t \nabla^{m+1} x_{t+k}; \\ \hat{r}_{y_k} &= \frac{1}{n-m-1} \sum_{t=m+2}^{n-k} \nabla^{m+1} y_t \nabla^{m+1} y_{t+k}; \\ \hat{r}_{x_k}' &= \frac{1}{n-m-1} \sum_{t=m+2}^{n-k} \nabla^{m+1} w_{x_t} \nabla^{m+1} w_{x_{t+k}}; \\ \hat{r}_{y_k}' &= \frac{1}{n-m-1} \sum_{t=m+2}^{n-k} \nabla^{m+1} w_{y_t} \nabla^{m+1} w_{y_{t+k}}; \end{aligned} \tag{22}$$

Simplification of Estimation Model of Digitizing Error in the Following Four Special Cases

(1) If the digitizing process is not stochastic or the random motion is too small to be considered, then Equation 20 may be reduced to

$$\hat{\sigma}_x^2 = \frac{1}{\lambda}\,\hat{\sigma}_{\mu_x}^2; \qquad \hat{\sigma}_y^2 = \frac{1}{\lambda}\,\hat{\sigma}_{\mu_y}^2. \tag{23}$$

(2) The line function $f(x, y) = 0$ is reduced to a point (a_0, b_0) if m is equal to zero in Equation 3. Then Equation 20 is reduced to

$$\hat{\sigma}_x^2 = \frac{1}{2(n-1)} \sum_{t=2}^{n} (x_t - x_{t-1})^2;$$
$$\hat{\sigma}_y^2 = \frac{1}{2(n-1)} \sum_{t=2}^{n} (y_t - y_{t-1})^2, \tag{24}$$

Equation 24 is an error estimation formula for repeatedly digitizing a point n times.

(3) The line function $f(x, y) = 0$ becomes a straight line $(\bar{x} - a_0)/a_1 = (\bar{y} - b_0)/b_1$ if m is equal to one in Equation 3. Then Equation 20 is reduced to

$$\hat{\sigma}_x^2 = \frac{1}{6(n-2)} \sum_{t=3}^{n} (x_t - 2x_{t-1} + x_{t-2})^2;$$
$$\hat{\sigma}_y^2 = \frac{1}{6(n-2)} \sum_{t=3}^{n} (y_t - 2y_{t-1} + y_{t-2})^2, \tag{25}$$

Equation 25 is an error estimation formula for digitizing a straight line.

(4) If $m = 2$ in Equation 3, the line function $f(x, y) = 0$ becomes a quadratic curve shown as follows:

$$\bar{x}_t = \sum_{k=0}^{2} a_k t^k; \qquad \bar{y}_t = \sum_{k=0}^{2} b_k t^k.$$

Then Equation 20 is reduced to

$$\hat{\sigma}_x^2 = \frac{1}{20(n-3)} \sum_{t=4}^{n} (x_t - 3x_{t-1} + 3x_{t-2} - x_{t-3})^2;$$
$$\hat{\sigma}_y^2 = \frac{1}{20(n-3)} \sum_{t=4}^{n} (y_t - 3y_{t-1} + 3y_{t-2} - y_{t-3})^2. \tag{26}$$

Determination of the Backward Difference Operator's Order ($d = m + 1$)

The efficiency of filtering the trend motion from a stochastic observation sequence depends on the selection of the backward difference operator's order $d = m + 1$. Hence, choosing an appropriate d is a key problem for generating a random field and for building an efficient estimation model of the digitizing error. A proper d could be identified by means of examining the autocovariance functions of the original observation sequence and a new stochastic sequence after having completed the backward difference process if a map line is short and simple. However, if the map line is long and complicated, then we should divide this line into several parts according to the type of trend and then determine order d based on the complexity of each part.

Amrhein and Griffith (1991) qualitatively classified different kinds of lines in terms of complexity of them (see Figure 1). Line (a) is the simplest and line (d) is the most complicated. For establishing the relationship between d and degree of complexity, it is necessary to derive some mathematical formulae which could be used to describe the complexity quantitatively.

Manual digitizing process is a kind of low speed discrete data acquisition. After having digitized a map line, a point sequence with known coordinates of knot points has been generated. In other words, this line is represented approximately by $(n - 1)$ broken lines.

The complexity of a line could be evaluated quantitatively by the following factors: (1) degree of tortuosity and (2) degree of fluctuation. The first factor is the total sum of absolute values of turning angles from the start point to the end point. The second one is a kind of global measure of deviation of knot points from the "reference line." The reference line is the standard line to be used for measuring the degree of fluctuation.

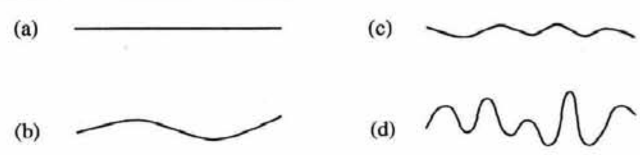

Figure 1. Lines with different complexities.

Definition of Degree of Tortuosity

In fact, the degree of tortuosity is not only determined by the sum of turning angles, but is also related to scale. In other words, the degree of tortuosity of a line is reduced if the scale of this line is enlarged when the sum of turning angles keeps unchanged. To avoid this problem, the definition of degree of tortuosity should be extended as follows:

$$\xi = \frac{\lambda \sum_{k=2}^{n-1} \alpha_k,}{\sum_{k=1}^{n-1} \|\vec{L}_k\|} \tag{27}$$

where λ is the scale factor, and α_k and $\|\vec{L}_k\|$ are the turning angle and length of line L_k between points k and $k + 1$. So if a map line could be regarded as one consisting of $(n - 1)$ vectors $\vec{L}_k$ $(k = 1,2, \cdots, n - 1)$, then we have

$$\alpha_k = \arccos\left(\frac{\vec{L}_k \cdot \vec{L}_{k+1}}{\|\vec{L}_k\| \|\vec{L}_{k+1}\|}\right), \; k = 2,3, \cdots, n = 1 \tag{28}$$

where $\vec{L}_k = (x_{k+1} - x_k)\,\vec{i} + (y_{k+1} - y_k)\,\vec{j}$, $\vec{i} = (1,0)$ and $\vec{j} = (0,1)$ are the unit vector of coordinates; $\|\vec{L}_k\| = [(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2]^{\frac{1}{2}}$; and $\vec{L}_k \cdot \vec{L}_{k+1} = (x_{k+1} - x_k)(x_{k+2} - x_{k+1}) + (y_{k+1} - y_k)(y_{k+2} - y_{k+1})$ are the scalar products. Apparently, the degree of tortuosity of a line can be computed by using Equation 27 if the coordinates of knot points are known.

Definition of Degree of Fluctuation

For defining the degree of fluctuation, we need to give the definition of the reference line and methods of measuring the degree of fluctuation.

Definition of median line operation ∇_m: assuming that a line is described by a point sequence $\{P(n) \mid P_k, k = 1,2, \cdots, n\}$, the middle point of line L_k is defined by $P_k^1 = med\{P_k, P_{k+1}\}$. The connection line between two adjacent middle points is called the median line. This results in a new broken line $L_m^1(u)$ consisting of a point sequence $\{P^1(u) \mid P_k^1, k = 1,2, \cdots, u\} \cdot L_m^1(u)$ is also called the first-order median line of the original line $L(n)$ and denoted by $L_m^1(u) = \nabla_m L(n)$. Similarly, the d-order median line operator is expressed by $L_m^d(u) = \nabla_m^d L(n)$. Obviously, the d-order median line operator satisfies $\nabla_m^d L(u) = \nabla_m^{d-1}(\nabla_m L(n))$, where $\nabla_m^0 \equiv 1$. The median line operator has two properties: (1) If line $L(n)$ is a closed broken line consisting of n points $\{P(n) \mid P_k, k = 1,2, \cdots, n\}$, then $L_m^d(u) = \nabla_m^d L(n)$ is still a closed broken line with the same number of knot points, $u = n$, but the area and perimeter of polygon $L_m^d(n)$ is smaller than $L_m^{d-1}(n)$. If $d \to \infty$, $L_m^d(n)$ will be reduced to a point (see Figure 2a); (2) If line $L(n)$ is an open convex broken line, after $d = (n - 1)$-order $\nabla_m^d L(n)$ operation, line $L(n)$ will be reduced to a straight line (see Figure 2b). This straight line is selected as the reference line in this paper. In the computation of degree of complexity, a closed broken line could be treated as an open broken line if any two adjacent knot points are assumed to be disconnected.

There are many methods of measuring the departure of

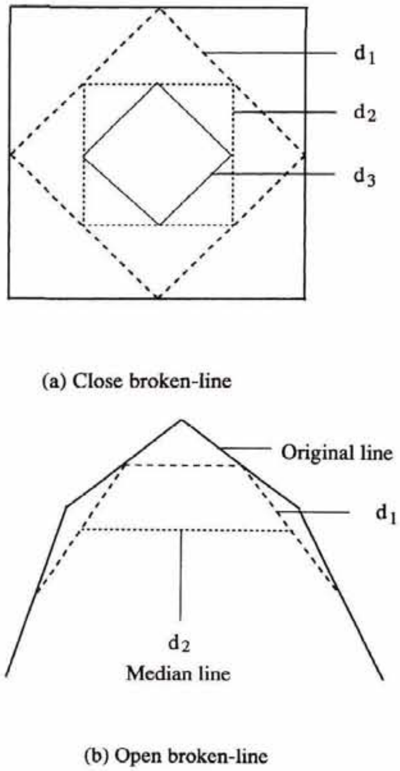

Figure 2. Broken lines and median lines.

knot points from the reference line; for instance, the sum of distances of knot points of a broken line from the reference line or the largest one of them. In this paper, the mean square distance is chosen as the total measure of fluctuation: i.e.,

$$\delta = \frac{1}{\lambda}\sqrt{\frac{\sum_{k=1}^{n} e_k^2}{n}}. \tag{29}$$

where e_k denotes the distance from knot point k to the reference line. Let's assume that $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_2, \bar{y}_2)$ are the coordinates of both end points of the reference line, respectively; then this reference line can be expressed by the following equation:

$$\begin{vmatrix} x & y & 1 \\ \bar{x}_1 & \bar{y}_1 & 1 \\ \bar{x}_2 & \bar{y}_2 & 1 \end{vmatrix} = 0, \tag{30}$$

which can be written in the form

$$(\bar{y}_1 - \bar{y}_2)\, x_k + (\bar{x}_2 - \bar{x}_1)\, y_k + (\bar{x}_1\bar{y}_2 - \bar{y}_1\bar{x}_2) = 0.$$

Distance e_k can be computed by

$$e_k = \frac{|(\bar{y}_1 - \bar{y}_2)\, x_k + (\bar{x}_2 - \bar{x}_1)\, y_k + (\bar{x}_1\bar{y}_2 - \bar{y}_1\bar{x}_2)|}{\sqrt{(\bar{y}_1 - \bar{y}_2)^2 + (\bar{x}_2 - \bar{x}_1)^2}}. \tag{31}$$

Definition of Degree of Complexity

With considering factors (1) and (2) simultaneously, the degree of complexity of line could be described by the following equation:

Table 1. Degree of Complexity of Four Broken-Lines

Lines	ξ	δ	γ
Straight line	0	0	0
Regular triangle	$\frac{\sqrt{2}}{18R}\pi$	$\frac{3}{4}R$	$\frac{1}{4}\pi R$
Square	$\frac{\sqrt{2}}{6R}\pi$	$\frac{\sqrt{10}}{4}R$	$\frac{\sqrt{10}}{4}\pi R$
Regular hexagon	$\frac{8}{15R}\pi$	$\frac{\sqrt{755}}{4}R$	$\frac{2\sqrt{755}}{3}\pi R$

Note: R is the radius of the circle as shown in Figure 3.

$$\gamma = \delta \sum_{k=2}^{n-1} \alpha_k. \tag{32}$$

For better understanding the concept of measuring degree of complexity, the degrees of complexity of four open broken lines: straight line, regular triangle, square, and regular hexagon are calculated by using Equation 32 and are listed in Table 1.

Numerical Examples

Efficiency of the Estimation Model of Digitizing Error Evaluated by Using Simulating Data

To examine the efficiency of the backward difference operators and approaches proposed for determining appropriate order of the operators, several sets of simulating data were generated. Three types of lines (33a), (33b), and (33c) were used in this simulation, where line (33a) is simple, line (33b) is complicated, and line (33c) is the most complicated: i.e.,

$$\begin{cases} x_t = 1 + t + t^2 \\ y_t = 1 + t + 2t^2 \end{cases} \quad (t = 1,2, \cdots, 100) \tag{33a}$$

$$\begin{cases} x_t = 1 + t + t^2 + t^3 \\ y_t = 1 + t + 2t^2 + 3t^3 \end{cases} \quad (t = 1,2, \cdots, 100) \tag{33b}$$

$$\begin{cases} x_t = 1 + t + t^2 + t^3 + t^4 \\ y_t = 1 + t + 2t^2 + 3t^3 + 4t^4 \end{cases} \quad (t = 1,2, \cdots, 100) \tag{33c}$$

Random motion can be generated by the following two-dimensional p-order autoregressive functions:

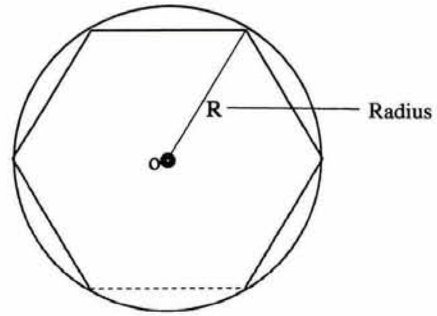

Figure 3. Regular hexagon with circumscribed circle.

Table 2. Effect of d and p on the Estimation of Digitizing Error

d	p	$\hat{\sigma}_x$ (mm)	$\hat{\sigma}_y$ (mm)	ζ
	0	±0.332	±0.329	0.9962
2	1	±0.173	±0.178	0.9990
	2	±0.130	±0.127	0.9994
	0	±0.101	±0.098	0.9990
3	1	±0.122	±0.123	0.9998
	2	±0.157	±0.160	0.9990
	0	±0.131	±0.129	0.9992
4	1	±0.182	±0.185	0.9987
	2	±0.323	±0.320	0.9954

Note: ζ denotes the degree of fitting.

$$p = 0\text{: no random motion} \tag{34a}$$

$$p = 1\text{: } \nabla^d z_t - \phi_1 \nabla^d z_{t-1} = \mu_t \tag{34b}$$

$$p = 2\text{: } \nabla^d z_t - \phi_1 \nabla^d z_{t-1} - \phi_2 \nabla^d z_{t-2} = \mu_t \tag{34c}$$

$$p = 3\text{: } \nabla^d z_t - \phi_1 \nabla^d z_{t-1} - \phi_2 \nabla^d z_{t-2} - \phi_3 \nabla^d z_{t-3} = \mu_t \tag{34d}$$

where

$$\phi_1 = \begin{pmatrix} 0.1 & 0.2 \\ 0.1 & 0.4 \end{pmatrix}, \phi_2 = \begin{pmatrix} 0.2 & 0.3 \\ 0.2 & 0.4 \end{pmatrix}, \phi_3 = \begin{pmatrix} 0.3 & 0.6 \\ 0.4 & 0.5 \end{pmatrix},$$

and $\{\mu_t\}$ is the white noise sequence with zero mean vector and variance 0.12mm $\times I_2$; here, I_2 is the two-dimensional unit matrix.

The simulating test was carried out in two steps:

(1) Line (33a) and autoregressive function (34b) were selected for generating a simulating data. In this case, $d = 3$, $p = 1$ could be considered as the most proper values of d for the backward difference operator and p for the autoregressive function, respectively, and $\hat{\sigma}_x = \pm$ 0.120mm, $\hat{\sigma}_y = \pm$ 0.120mm could be regarded as theoretical values of digitizing error in this simulating data.

"Models" consisting of different combinations of $d = 2,3,4$ and $p = 0,1,2$ were employed to fit the simulating data and to examine what happens if the orders are chosen improperly. The results of this experiment are listed in Table 2.

It is clear to see from Table 2 that model with $d = 3$, $p = 1$ has the best degree of fitting ($\zeta = 0.9998$) and estimated values $\hat{\sigma}_x = \pm$ 0.122mm and $\hat{\sigma}_y = \pm$ 0.123mm are the closest to their theoretical values.

(2) The second experiment is to use lines (33a), (33b), (33c), and autoregressive function (34b) to generate three sets of simulating data. The degrees of complexity of these three lines could be computed by using the simulating data based on Equation 32. The values of d and p are chosen depending on the calculated degrees. The computed results are listed in Table 3.

It is shown apparently in Table 3 that the estimated values of digitizing error are very close to their theoretical values. In other words, the estimation model may not be sensitive to the type of line being digitized if the value of d is selected appropriately. Because the length of this paper is limited, the problem of how to choose a proper value of d in terms of degree of complexity will be discussed in details in our future paper, "Approaches for Separating Trend Motion from Stochastic Sequence Series for Building Estimation Model of Digitizing Error."

Efficiency of the Estimation Model of Digitizing Error Evaluated by Using Real Data

The efficiency of the estimation model of digitizing error proposed in this paper has been evaluated theoretically. It is necessary to examine the efficiency by using real data. A circle, a contour line, and a seacoast line on a 1:10,000-scale topographic map have been digitized by using a Calcomp 9100 digitizer. The sizes of these three sets of digitized data are 65 by 2, 87 by 2, and 183 by 2, respectively. The computed results are listed in Table 4.

Table 3. Evaluation of the Estimation Models Using Simulating Data

Line	d	γ	$\hat{\sigma}_x$ (mm)	$\hat{\sigma}_y$ (mm)
33a	2	1.21	±0.122	±0.123
33b	3	3.08	±0.120	±0.124
33c	4	5.37	±0.125	±0.126

Table 4. Evaluation of the Estimation Models Using Real Data

Line	d	p	$\hat{\sigma}_x$ (mm)	$\hat{\sigma}_y$ (mm)	ζ
Circle	3	1	±0.135	±0.137	0.9996
Contour	3	1	±0.132	±0.134	0.9996
Seacoast	3	0	±0.131	±0.130	0.9997

The specifications for the Calcomp 9100 digitizer are: dimension: 36 by 48'; resolution: 0.001'; digitizing accuracy: 0.010 ± 0.0005'; and the maximum positioning error ≤ 0.020'.

Table 4 shows that the estimated values of digitizing error are slightly different. It indicates that the estimation model proposed in this paper may not be sensitive to the type of line to be digitized. The small difference in the results could be explained because digitizing a complicated line will lead to more difficulty for an operator to control a cursor in positioning than would digitizing a simple line.

Conclusion

Theoretical estimation models of digitizing error have been derived in this paper. In addition, the models have been simplified for practical use. The key problem in this paper is to search for ways to be used for efficiently removing the trend motion from a stochastic sequence series of digitizing data. For a short and simple map line, the backward difference process could be an efficient approach to filter the trend motion from the random sequence series. For a long and complicated line, however, it could be impossible to find a polynomial to simulate that line. In this case, this line should be divided into several parts based on the type of trend. To improve the efficiency of the backward difference process, the order of the backward difference operators should be selected in terms of the type of trend in each part. Because the lines on the map are usually complicated, they are not easily expressed by polynomial functions mathematically, From this point of view, the degree of complexity of line suggested in this paper could be an efficient way to describe the type of the line in general. For choosing the value of order of the backward difference operator appropriately, several approaches for measuring the degree of complexity of line have been proposed.

References

Amrhein, C.G., and D.A. Griffith, 1991. *A Statistical Model for Analyzing Error in Geographic Data in an Information System*, Discussion paper No. 38, Dept. of Geography, University of Toronto.

Box, G.E.P., and G.M. Jenkins, 1976. *Time Series Analysis: Forecasting and Control*, Holden-Day.

Burrough, P.A., 1986. *Principle of Geographic Information Systems for Land Resources Assessment*, Clarendon Press, Oxford.

Caspary, W., and R. Scheuring, 1993. Positional accuracy in spatial databases, *Comput. Environ., and Urban Systems*, 17:103–110.

Chrisman, N.R., and B. Yandell, 1988a. A model for the variance in area, *Surveying and Mapping*, 48:241–246.

———, 1988b. Effects of point error on for area calculations: a statistical model, *Surveying and Mapping*, 48(4):241–246.

Goodchild, M.F., and O. Dubuc, 1987. A model of error for choropleth maps, with application to geographic information systems, *Proceedings, AutoCarto 8*, ASPRS/ACSM, Falls Church, Virginia, pp. 167–174.

Goodchild, M.F., and S. Gopal, 1992. *The Accuracy of Spatial Databases*, Taylor & Francis, London, pp. 107–108.

Graferend, E.W., and F. Sanso, 1985. *Optimization and Design of Geodetic Networks*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 588 p.

Haining, R.P., D.A. Griffith, and R.J. Bennett, 1983. Simulating two-dimensional autocorrelated surfaces, *Geographic Analysis*, 15: 247–255.

Herzog, A., 1992. Modeling reliability on statistical surfaces by polygon filtering, *The Accuracy of Spatial Databases* (M.F. Goodchild and S. Gopal, editors), Taylor & Francis, London, 18. pp. 209–216.

Keefer, B.J., J.L. Smith, and T.G. Gregoire, 1988. Simulating manual digitizing error with statistical models, *Proceedings, GIS/LIS'88*, ASPRS/ACSM, Falls Church, Virginia, pp. 475–483.

Lavenda, B.H., and C. Scherer, 1987. The statistical inference approach to generalized thermodynamics: I. Statistics; II. Thermodynamics, *Nuovo Cimento*, 100B:199–227.

Maffini, G., M. Arno, and W. Bitterlich, 1992. Observations and comments on the generation and treatment of error in digital GIS data, *The Accuracy of Spatial Databases* (M.F. Goodchild and S. Gopal, editors), Taylor & Francis, London, pp. 56–67.

Priestley, M.B., 1981. *Spectral Analysis and Time Series*, Academic Press.

Tobler, W.R., 1992. Frame independent Spatial Analysis, *The Accuracy of Spatial Databases* (M.F. Goodchild and S. Gopal, editors), Taylor & Francis, London, pp. 115–121.

Walker, G., 1931. On periodicity in series of related terms, *Pro. Royal Soc.*, A131:518.

Yule, J.W., 1927. On a method of investigating periodicity in disturbed series, with special reference to Wolfer's sunpot numbers, *Phil. Trans.*, A226:267.

(Received 26 April 1996; revised and accepted 10 March 1997)

ASPRS-RTI

A NEW VISION

Special Events planned include:

- Golf Tournament
- Busch Gardens Theme Park

1998 ASPRS-RTI Annual Conference
March 30–April 3, 1998
Tampa Convention Center
Tampa, Florida

"A New Vision" not only describes our industry at a time when new sensors and technologies are emerging at a rapid pace, but also captures our enthusiasm as ASPRS unveils its new conference look and feel.

CONFERENCE SCHEDULE

Committee Meetings
Monday–Wednesday
March 30–April 1

Full-Day Workshops
Monday–Tuesday
March 30–March 31

Educational Sessions and Exhibits
Wednesday–Friday
April 1–April 3

FULL REGISTRATION FEES

	Member	Non-Member
Early-Bird *(deadline January 15)*	$250	$350
Advance *(dealine March 1)*	$310	$410
On-Site	$350	$450

Preliminary Programs will automatically be mailed to ASPRS and RTI members in November. If you would like to be added to the mailing list, contact ASPRS at **301-493-0290 x20** or e-mail us at **meetings@asprs.org.**

SEE YOU IN SUNNY TAMPA, FLORIDA!

For updates and information, check out www.asprs.org/asprs.

1934 American Society for Photogrammetry and Remote Sensing

Resource Technology