Linear Feature Modeling with Curve Fitting: Parametric Polynomial Techniques

Xiaoming Zheng and Peng Gong

Center for Assessment and Monitoring of Forest and Environmental Resources
Department of Environmental Science, Policy and Management,
University of California, Berkeley, CA 94720-3310, USA

Abstract

A decomposition model is described to model linear features sampled by manual digitization or field survey. The model consists of three components, original data, systematic pattern, and random error. Least squares and moving least squares techniques are introduced for polynomial curve fitting. Polynomial functions are proposed to represent linear features. The position deviation between sampled points and the polynomial function is used as an approximation of the random error. Experimental results are presented to show the effectiveness of the decomposition model. Potential applications of the model have been discussed including estimation of errors associated with points sampled along linear features, digital representation and mapping of linear features.

摘 要

本文介绍对手工数字化或野外测量的线性特征进行模拟的方法。本模拟有三个组成部分：模拟出的原始数据、系统模式和随机误差。我们的模型使用多项式参数曲线拟合技术配以最小二乘法和滑动最小二乘法技术进行参数估算。所得的多项式参数函数可用来模拟曲线。由多项式函数和数字化点之间的偏差用来近似估算随机误差。试验结果表明我们的模型有一定功效。该模型在数据误差估算、线性特征的模式、数字表达和线性特征制图等有应用潜力。

I. INTRODUCTION

In a vector-based GIS, digital representation of curve features is done through the use of a series of point coordinates sampled along the curves. Manually digitizing paper maps is a predominant method of point sampling. This has been recognized as a significant source of error of spatial data (Chrisman, 1982). Perhaps a more precise method is to use global positioning systems (GPS) units in the field. The high precision of GPS receivers, however, does not offer much help for curve features because such features are approximated by and handled as a series of straight line segments joining the consecutively sampled points. The facts that use of discrete points to represent curve features is prone to errors and that such errors are not quantified are fundamental problems, some yet unsolved tasks, in vector-based GIS (Brunsdon and Openshaw, 1993).

To exactly calculate position errors of discrete points, we need to compare the sampled points with their true position along a curve. In most cases, however, the true curve is not known. We must first approximate the true curve from sampled points and then derive position uncertainties by comparing the sampled points with the estimated curve. To do so, we must (1) develop a mathematical model that can approximate and represent the true curve features in a map or in reality based on the sampled points, and (2) provide a procedure to estimate the errors or uncertainties associated with the curve model. Splines function used for curve fitting and interpolation is not suitable for those purposes because it forces the fitted curve to go through the sampled points. In time series analysis, some prediction models such as the autoregressive model, the Autoregressive Integrated Moving Average (ARIMA) model and adaptive filtering based on Wiener-Levinson and Kalman Filter theories may be useful, but they use data observed in the past to predict the future behavior of the modeled phenomenon (Janacek and Swift, 1993, Graupe, 1984). In spatial data modeling, it is desirable to use both the "past" and "future" points. Some shape

1082-4006/97/0301~2-7$3.00
©1997 The Association of Chinese Professionals in
Geographic Information Systems (Abroad)

analysis methods (e.g., Lin and Hwang, 1987; Gunther and Wong, 1990; Grogan et al., 1992) including strip tree, curve fitting method using Bezier curves, arc tree, and Fourier descriptors may only be useful for curve representation not for estimation of errors or uncertainties.

Polynomials can be used for the two purposes mentioned above, particularly for intuitively smooth curves that are continuously differentiable. In reality, not all linear features have this mathematical property. Many linear features resulting from human activities may not be continuously differentiable. Examples are roads, utility lines, cadastral and administration lines. On the other hand, most linear features delineating natural phenomena can be considered as continuously differentiable. These include contours, streams, and natural resource boundaries (e.g., soil, climate, vegetation, wetland, etc.). Because of the increased amount of human abstraction realized by map generalization, on smaller scale maps we observe a larger proportion of differentiable curves. It is possible to store polynomial coefficients and use polynomial functions to represent curve features particularly if lower order polynomial functions can fit curves with sufficiently high accuracy. It may require less space to store polynomial coefficients than to store sampled point coordinates. In addition, it is effective to use polynomial coefficients to represent curve shapes. Curve shape analysis may be made based on polynomial functions for subsequent curve generalization, curve matching for object registration or recognition. Thus, curve representation with polynomial coefficients may have some advantages in data storage and curve shape analysis over the traditional curve representation method involving consecutive straight lines.

The objective of this paper is to develop a decomposition model for curve fitting by employing polynomial functions. The model consists of three components, original data, systematic error, and random error. It is used to simulate differentiable curve features from sampled points and to approximate sampling errors or uncertainties. Without loss of generality, we concentrate on the development of the model and its application to digitized curve features. Sampled points through GPS units can be processed in the same manner. In the next section, we introduce a framework for spatial data modeling particularly curve modeling based on the decomposition model. In section 3, we introduce an epsilon band model for the estimation of errors or uncertainties of sampled points that constitute a stationary random data series. For estimation of polynomial coefficients, we describe least squares and moving least squares methods in sections 4 and 5, respectively. The two methods are used to implement the curve models. Some experiment results with digitized data from simulated curves are presented in section 6 followed by some conclusions.

II. SPATIAL DATA MODELING

There are two kinds of natural or social phenomenon that can be described with a mathematical model. One is deterministic physical process or signal, which is entirely known and can be represented exactly with a mathematical function. The other is random event, which can only be described using a stochastic model based on random samples.

Spatial data digitization is a stochastic process (Keefer et al, 1988). The random error is introduced during the generation, analysis and processing of the digitized spatial data. To model a digitized line, its uncertainty and random error, three steps are needed. These are model selection, model estimation, and model evaluation.

Model Selection

A spatial series model describing a curve or a linear feature should be capable of (1) representing the original data with a deterministic mathematical function that can simulate or account for the sampled data series, and (2) estimating the random error distribution for evaluating the accuracy of the fitting, interpolation and prediction of the linear feature. Selecting an appropriate model is one of the most crucial steps in spatial data modeling. Model selection criteria depend on the objective of data simulation. To describe the behavior of a physical phenomenon, we may derive a model based on physical laws so that we can precisely represent or predict the value of a physical parameter in a given time or space. To estimate the trend of a random phenomenon affected by many unknown factors, stochastic process models can be chosen. It is helpful to plot the data first for understanding the type, pattern and trend of the random data sets. The knowledge on the physical process of data acquisition is also useful in mathematical model selection.

For a discrete series represented by digitized points {P_t, t = 1, 2, ... , n}, the series can be expressed with a general stochastic decomposition model

$$P_t = F_t + E_t + R_t \qquad (1)$$

where

- P_t is the digitized point;
- t is a number index for a particular point;
- F_t is the component that represents the original, undistorted part of the data series;
- E_t contains the systematic pattern or systematic error that can be removed if known.
- R_t is the random component, which can only be estimated using some a priori knowledge about its distribution.

A digitized-point series consists of a sequence of X and Y coordinates, which can generally be represented as:

$$\{P_t = (X_t, Y_t), \quad t = 1, 2, \cdots, n\} \qquad (2)$$

Because sample points are a discrete series with unequal intervals, it is more practical to write the point set in a parametric form

$$\begin{cases} P_t = (X_t, Y_t) \\ X_t = X(s_t) \\ Y_t = Y(s_t) \\ t = 1, 2, \cdots, n \end{cases} \qquad (3)$$

where s_t is a distance parameter between the origin and point t. Both X_t and Y_t are the functions of parameter s_t . X_t and Y_t can be fitted by using the same mathematical form with different coefficients. We only focus on the discussion of data modeling with the series of $\{X_t, \quad t = 1, 2, \cdots, n\}$in this paper. The decomposition model can be written as

$$X_t = f_t + e_t + r_t \qquad (4)$$

The component of e_t largely depends on the physical process of data acquisition and the digitizer used. It is difficult to use a mathematical expression to describe the systematic pattern without a complete knowledge of a specific data series to be modeled. One of the common systematic errors is linear shift. To remove the systematic effect of a linear shift, a linear parametric function can be used to rectify the error. Some systematic patterns or errors may be modeled separately through visual analysis of the digitized data. Visualizing the sample points may allow systematic errors to be detected and corrected through manual editing. To simplify the discussion, we assume that there is no systematic pattern and error in a digitized data series, that is, the component of e_t is zero. We have

$$X_t = f_t + r_t \qquad (5)$$

Generally, any data series can be decomposed into a deterministic and a random part and can be represented by equation (5) (Janacek and Swift, 1993). Model selection includes the determination of the mathematical expression for the parametric equation f_t and the definition of the distribution of the random component r_t.

Model Estimation

A mathematical model for describing a random event contains some unknown parameters, which should be estimated with the available sample data. Least squares is an important statistical technique for estimating model parameters based on some specified standard and criteria.

Suppose a model for a spatial data series takes the form

$$X_t = f_t + r_t = f(t,\theta) + r_t \qquad (6)$$

where θ isthe parameter vector and $\theta = (\theta_1, \theta_2, ..., \theta_k)^T$.

Let $\hat{X}_t = f(t,\theta)$ be an estimation of the original data set X_t'. Then the deviation of the estimation for the t-th data point is

$$r_t = X_t - \hat{X}_t, \qquad t = 1, 2, ..., n \qquad (7)$$

The sum of squared deviations is

$$R = \sum_{t=1}^{n} r_t^2 = \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 \qquad (8)$$

The criterion for calculating parameter θ_i is that the parameter can minimize the sum of the squares. Let

$$\frac{\partial R}{\partial \theta_i} = 0, \qquad i = 1, 2, ..., k \qquad (9)$$

Since $\hat{X}_t$ is a linear function of the parameter vector θ, we can draw a set of k linear equations from (9). Solving the linear equation set, we can obtain the parameters, θ_i (i=1, 2, ..., k; k≤n), which minimize the sum of squared deviations in (8).

Model Evaluation

When a model is estimated based on the available data sets, it is necessary to diagnostically check the goodness of fit between the estimated and the digitized data. We need to ascertain if the model is appropriate for the data set, and evaluate the estimated characteristic parameters.

III. STATIONARY RANDOM DISTRIBUTION AND EPSILON MODEL

Suppose there is a random data series $\{X_t | t = 1, 2, \dots\}$. If its statistical characteristics do not change with variable t, that is, the characteristics is independent of the origin of variable t, we call the random data series stationary (Janacek and Swift, 1993).

A stationary data series have a constant mean

$$\mu_t = E[X_t] = \mu \qquad (10)$$

and, for any two points t and s in time series, its autocovariance function satisfies

$$R(s-t) = E[(X_s - \mu)(X_t - \mu)] \qquad (11)$$

Let m=s-t , and we have

$$R(m) = E[(X_{t+m} - \mu)(X_t - \mu)] \qquad (12)$$

Specially, if m=0, the autocovariance becomes the squared deviation

$$R(0) = E[(X_t - \mu)^2] \qquad (13)$$

A stationary random series is completely characterized by its mean and autocovariance. The exact values of these parameters can be calculated if the ensemble of all possible realizations is known. Otherwise, they can be estimated if multiple independent realizations are available. However, in most applications, it is difficult or impossible to obtain multiple realizations. Most available spatial data series constitute only a single realization. This makes it impossible to calculate the ensemble average. For a stationary data series, we have a natural alternative of replacing the ensemble average by the average along the time or distance axle if the stationary process is ergodic (Zhong and Hu, 1990).

The process of digitization can be considered as a stationary random sampling process when the cursor is used to trace a curve which can be modeled by a normal distribution (Keefer et al, 1988; Maffini et al 1989; Gong and Chen, 1992). This implies that the probability of the sampled points located at both sides of the curve are about the same and the sum of all the errors cancels out. However, for a random sample point, it is impossible to predict its position along the curve. Moreover, the true curve is usually unknown. It has been suggested that the epsilon band model proposed by Perkal (1966) be used to represent an uncertain zone centered at the continuous representation of the digitized curve (e.g., Blakemore, 1984). The position deviation, epsilon, times a certain number is used as the width of the uncertain zone (Figure 1).

In practice, the problem with applying the epsilon band model to indicate curve uncertainty is how to estimate the position deviation - the epsilon value. Distances between the representation of the curve and the digitized points (Bolstad et al, 1990; Gong and Chen, 1992) may be used to estimate the width of the epsilon band. Provided that the true curve can be simulated with a mathematical function, the digitizing error can be estimated from the sample

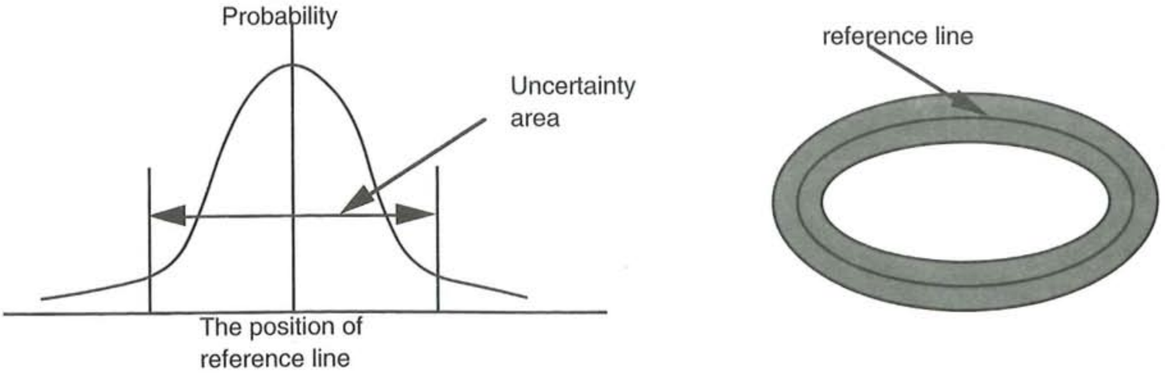


Figure 1. The distribution of digitizing points along a true boundary line and Epsilon band.

standard deviation of the stationary random process based on a series of digitized points.

$$\hat{r} = \sqrt{\hat{R}(0)} = \left[\frac{1}{n}\sum_{t=1}^{n}(X_t - f_t)^2\right]^{\frac{1}{2}} \qquad (14)$$

where f_t is supposed to be the points on the true curve, which is consistent with equation (5). Because f_t is unknown, a polynomial function can be used to represent f_t based on the available sample points.

IV. UNCERTAINTY MODELING WITH POLYNOMIAL CURVE FITTING

From equation (5), we have

$X_t = f_t + r_t$

where f_t represent the undistorted curve, and r_t is the random component with a normal stationary distribution.

There are two approaches to estimating the random error r_t, filtering and curve fitting.

The first approach is to use a filter to remove f_t from X_t , that is

$$\hat{r}_t = F\{X_t\} \qquad (15)$$

If the original curve is continuous and smooth, f_t should be a low frequency signal, which can be fitted by using a polynomial function. The component of r_t is mainly a high frequency signal, which usually has a normal distribution. To remove f_t, filtering can be applied in spatial or frequency domain.

In spatial domain, a high-pass filter is equivalent to taking the derivatives. If f_t can be represented by a K-th order polynomial function, a (K+1)th-order differential operator can remove f_t from X_t . The standard deviation of the differentiated result can be used as an estimation of the random error. The problem is that, for a discrete spatial data series, a difference operator has to be used to replace the differential operation. Difference operation is not invertible, and the operation will enhance the random component when removing the low frequency signal, which will change the magnitude of the error and affect the estimation of the parameters. High-pass filtering in frequency domain seems to be more reasonable for removing f_t and estimating the random error if a filter can be designed to remove the low-frequency part and keep the high-frequency part unchanged.

The second approach is to use a mathematical function to simulate f_t, and then subtract the estimated $\hat{f}_t$ from X_t, that is

$$\hat{r}_t = X_t - \hat{f}_t \qquad (16)$$

We use polynomials to fit a series of digitized points as a simulation of the true curve. The general form of a polynomial function is defined as

$$X = X(s) = \sum_{k=0}^{K} a_k s^k = a_0 + a_1 s + a_2 s^2 + ... + a_k s^k \qquad (17)$$

where K is a non-negative integer, the degree of the function, and $a_0, a_1, \cdots, a_k$ are fixed real numbers, s is the distance between point s and the origin which can be defined as the first digitizing point. The coefficients $a_0, a_1, \cdots, a_k$ can be calculated using the least squares technique based on the available sampled points.

Suppose that there are n digitized points for a curve. For the t-th point, the fitting equation is

$$\hat{X}_t = X(s_t) = \sum_{k=0}^{K} a_k s_t^k = a_0 + a_1 s_t + a_2 s_t^2 + ... + a_k s_t^k, \quad t=1,2,...,n \qquad (18)$$

The residue is

$$R = \sum_{t=1}^{n} r_t^2 = \sum_{t=1}^{n}(X_t - \hat{X}_t)^2 \qquad (19)$$

Let

$$\frac{\partial R}{\partial a_i} = 0, \qquad i = 0,1,...,k\,(k \le n) \qquad (20)$$

we have a set of k linear equations for $a_0, a_1, \cdots, a_k$. Solving these equations we can obtain the unique solution for all the coefficients. In practice, the shape of a curve to be fitted needs to be smooth and continuous and the order of the polynomials cannot be infinitely high. These are further elaborated below.

(1). Curve continuity

A curve is mathematically continuous and smooth if its various order of derivatives exist. If a curve is discontinuous or there exist sharp turning points, it should be divided into continuous and smooth segments at the broken points, and piecewise polynomial functions may be constructed segment by segment. Practically, polynomial functions are less effective for linear features that are intrinsically non-smooth or mathematically discontinuous because more sample points may be required and a curve may have to be broken into too many segments.

(2). The order of the polynomial function

Theoretically, a polynomial function can fit any continuous and smooth curve so long as the order is sufficiently high. Usually, better results can be obtained when a higher order of polynomial function is used to fit a curve if there is a sufficient number of sample points. Practically, there is a computational problem related to the limited precision and magnitude of a computer. An exceedingly high order will cause the fitted curve vibrating around the curve because of the intrinsic ill-condition of the Vandermonde problem and the roundoff errors, which may introduce rather substantial coefficients in the leading terms of the polynomial. A reasonable order for fitting a specific curve needs to be determined.

If a k-th order polynomial function can completely represent a curve, then (k+1)th derivative operation will result in zero. This fact can be used to construct a method to determine the order of a polynomial function.

For discrete sample points, the backward difference operator can be defined as (Wei, 1990):

$$\nabla X_t = X_t - X_{t-1} = (1-B)X_t \qquad (21)$$

and

$$\nabla^k X_t = (1-B)^k X_t \qquad (22)$$

where $\nabla = 1 - B$ and B is a backward shift operator $BX_t = X_{t-1}$.

After the k-th order difference operation, we need to test the assumption of random stationary distribution for the residue (Janacek and Swift, 1993). If the assumption is true, we take k as the appropriate order for the polynomial function. If a curve can be completely represented with a K-th order polynomial function, for any P-th order polynomial function (P>K), all the coefficients $a_j \approx 0$ $(K < j \le P)$, and the accuracy of the fitting should be the same as the K-th order polynomial function. There is a more practical method to determine the order of a polynomial curve. When k is smaller than K, the fitting error or residual R_k monotonically decreases as the order increases. When k increases to K+1 and if R_k equals R_{K+1} or even is less than R_{K+1}because of the intrinsic ill-condition of the Vandermonde problem and the roundoff errors, K should be taken as the order of the polynomials.

V. MOVING LEAST SQUARES FOR CURVE FITTING

If a continuous curve changes sharply in some parts and changes gently in some other parts, it requires a high order polynomial function to fit the curve. On the other hand, because the precision limitation of a computer, a sufficiently high order of polynomial function will cause vibration of the fitted curve and hence introduce a large fitting error. To solve this problem, moving least squares can be used.

The basic idea of moving least squares is that if X is the function associated with the fitted curve, then the value of X at a point s should be most strongly influenced by the values at those points s_t that are close to s. In other words, the influence of a value at s_t on X at point s should decrease as the distance between s and s_t increases. Therefore, we can modify equation (19) to a weighted sum of squared deviations and minimize

$$R_m = \sum_{t=1}^{n} w_t(s)[X(s_t) - X_t]^2$$

$$= \sum_{t=1}^{n} w_t(s)\left[\sum_{k=0}^{K} a_k s_t^k - X_t\right]^2 \qquad (23)$$

where $w_t(s)$ is the weight function of s at point t. Choose the following function

$$w_t(s) = \exp\left(-C(s - s_t)^2\right) \qquad (24)$$

which is a monotonically decreasing function, and C is a constant for adjusting weights of neighboring points. The greater C is, the smaller is the size of the neighborhood points that have significant effect on the fitted value at position s.

To obtain the optimal coefficients according to the minimum squares of deviation, let

$$\frac{\partial R_m}{\partial a_k} = 0, \qquad k= 0,\ 1,\ \dots,\ K \qquad (25)$$

Then, the normal equations are

$$
\begin{aligned}
& a_0\left(\sum_{t=1}^{n} w_t s_t^0\right) + a_1\left(\sum_{t=1}^{n} w_t s_t^1\right) + \cdots + a_K\left(\sum_{t=1}^{n} w_t s_t^K\right) \\
& \quad = \sum_{t=1}^{n} w_t X_t \\
& a_0\left(\sum_{t=1}^{n} w_t s_t^1\right) + a_1\left(\sum_{t=1}^{n} w_t s_t^2\right) + \cdots + a_K\left(\sum_{t=1}^{n} w_t s_t^{K+1}\right) \\
& \quad = \sum_{t=1}^{n} w_t s_t^1 X_t \\
& \quad \cdots \qquad \cdots \qquad \cdots \qquad \cdots \\
& a_0\left(\sum_{t=1}^{n} w_t s_t^K\right) + a_1\left(\sum_{t=1}^{n} w_t s_t^{K+1}\right) + \cdots + a_K\left(\sum_{t=1}^{n} w_t s_t^{2K}\right) \\
& \quad = \sum_{t=1}^{n} w_t s_t^K X_t
\end{aligned}
\tag{26}
$$

For polynomial curve fitting based on least squares, the coefficients of the polynomial function is identical for any fitted point s. Therefore, a high order polynomial function may be required. With moving least squares, the solution for the coefficients $a_0, a_1, \cdots, a_k$ depends on s through the weight function $w_t(s)$. For each s, we have to solve a set of normal equations. Therefore, it is computationally prohibitive to use high order polynomials.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

Experiments were conducted to evaluate the accuracy and uncertainty of digitization and curve fitting for linear features using polynomial functions based on least squares and moving least squares. The original curves were generated with some mathematical functions. The exact position error of each point can be calculated by comparing the digitized point, the fitted curve value with the original mathematical function.

Experimental procedures are as following:

Step 1. Design a kind of mathematical function to generate a curve.
Step 2. Print out the curve and digitize the curve using a digitizer.
Step 3. Fit the curve using a polynomial function based on least squares and moving least squares
Step 4. Estimate the random error and evaluate the accuracy of curve fitting.

Curve Generation

Different mathematical functions were used to generate different shapes of curves. To evaluate the effect of curvature on digitization, circles with different radii were used. It is more difficult to fit a curve that has different curvatures in its different segments. A set of sine functions were used to represent the curves with different curvatures. The sine function used to generate curves was $y = a\sin bx$, where a and b were parameters for adjusting the shape of the curve a={1.0, 1.25, 1.5, 2.0} and b={1.0} in our experiments. The third kind of curve was an ellipse representing polygon boundaries with different curvatures. The ellipse equation is:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

where a={2, 6} and b={1, 3}.

To further test the polynomial curve fitting technique, more complicated curves were constructed by mirroring a sine curve (Figure 6) or joining two ellipses (Figure 7).

Digitization

All the curves generated with mathematical functions were digitized manually by some experienced operators at their normal speeds. The digitized data were then taken as sampled points for curve fitting and random error estimation.

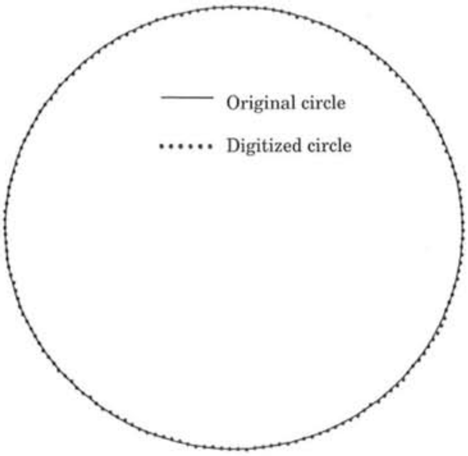


Figure 2. A digitized version of a circle (dotted) overlaid on top of the original one (solid).

Figure 2 shows a set of digitized points along one of the original circles. The radius of the circle is 5.5 cm printed with a laser printer having a resolution of 600 dots per inch (DPI). The line thickness is 0.1 mm. The digitization was done on a Summagraphics (MM II 1812) digitizer with a resolution of 1000 DPI. As can be seen from Figure 2, the digitized points are not exactly on the circle. We calculated the exact position errors and plotted the distribution of the digitizing errors in Figure 3. As expected, most of the digitized points locate along the circle and the number of sampled points decreases as the distance between the digitized points from the circle increases. Although the distribution of errors is a little skewed, it looks close to a normal distribution. In this study we assume that the digitizing error distribution is normal.

Errors caused from curve plotting by a printer and point-position reading from digitizing tables are determined by the resolution of the printer and digitizer used. Since a 600 DPI printer was used, curve plotting errors should be within approximately +/- 0.022 mm while point reading errors should be within 0.013 mm. Because errors from different sources do not simply add up (Gong et al., 1995) and the error magnitudes of curve printing and point reading are one order of magnitude less than the digitizing errors at an average level of approximately 0.2 mm (Figure 3), errors caused by curve printing and point reading have been ignored in this study.

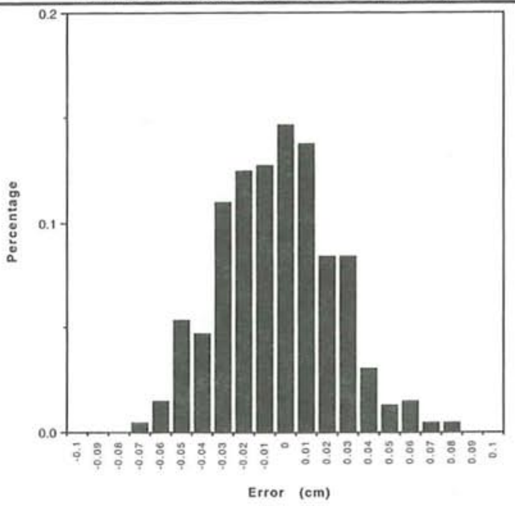


Figure 3. Digitized error distribution calculated from the example in Figure 2. Error is determined by calculating the distance of each digitized circle

Curve fitting

The curvature of the sine curve reaches its maximum at the peak and valley positions (Figure 4). With a 9-th order polynomial function, the curve fitting errors are still largely observable. Particularly, the fitted curve does not reach the apices of the sine curve (Figure 4a). Better results were achieved from the moving least squares with an order of 5 (Figure 4b).

Figure 5 shows a comparison of the results from the least squares and the moving least squares. The four curves in Figure 5(a) include the original curve and the curves generated by polynomial functions of order 1, 5, and 9, respectively. It can be seen that the curve simulated with the 9th order polynomials is a close approximation to the original curve. Figure

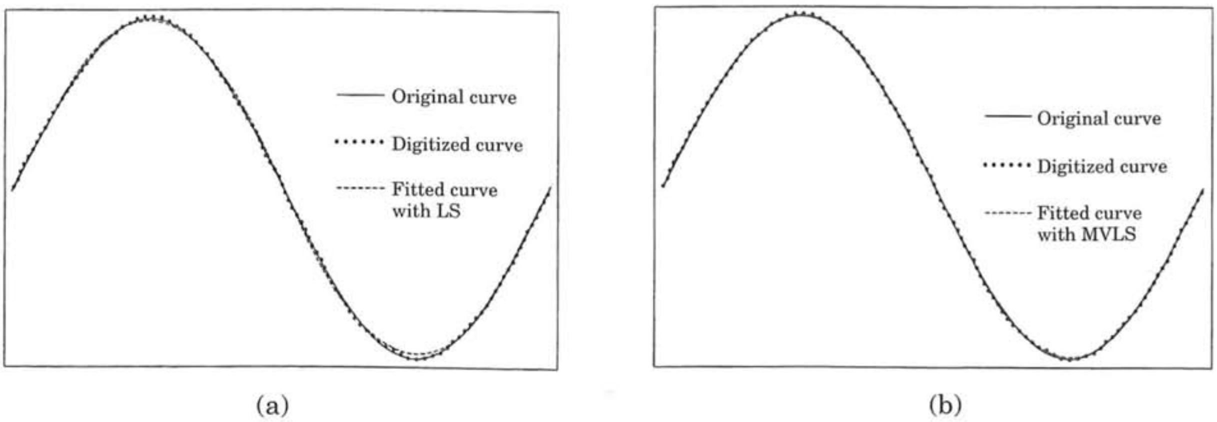


Figure 4. The effect of curvature on curve fitting. (a) Original curve, digitized version, and fitted curve using the 9th order polynomial functions estimated with the least squares method. (b) Fitted curve using the 5th order polynomials estimated with moving least squares.

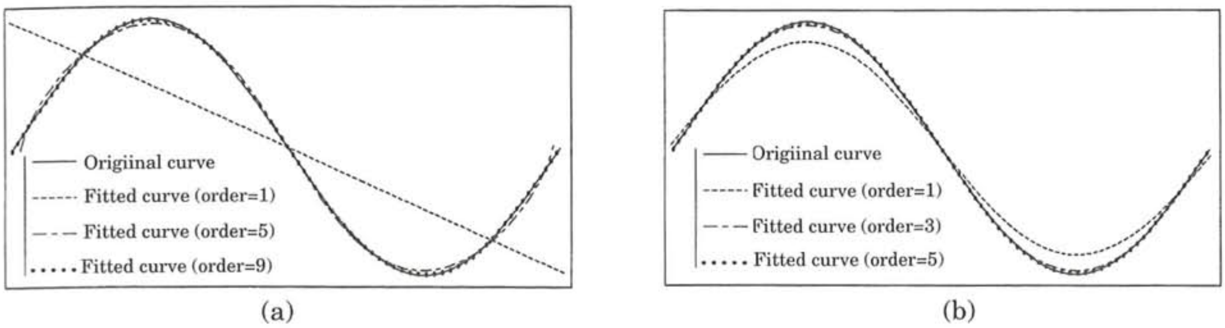


Figure 5. Curve fitting results using different orders of polynomial functions. The original is displayed using a solid line. (a) Fitted results using least squares with polynomial orders of 1, 5, and 9, respectively. (b) Fitted results using moving least squares with polynomial orders of 1, 3, and 5, respectively.

Table 1. The Comparison of the Errors of Least Squares and Moving Least Squares

Least Squares			Moving Least Squares		
Order	x (cm)	y (cm)	Order	x (cm)	y (cm)
1	0.463824	0.477288	1	0.029826	0.102721
3	0.059176	0.075351	2	0.011229	0.018969
5	0.023897	0.028138	3	0.008875	0.014013
7	0.006212	0.020102	4	0.004368	0.006309
9	0.006224	0.006068	5	0.003307	0.004847

5(b) shows the curve fitting results obtained from the moving least squares with order 1, 3, and 5, respectively. The curves generated with the 5th order polynomial functions fit well to the original curve. Table 1 summarizes some of the curve fitting accuracies. For the least squares method, when the order is 11, the sample variances along both the x and y directions are tremendously greater than those obtained from the 9th order polynomial functions. Therefore, an 11th order polynomial function may represent an over fitting to the original curve because of the intrinsic ill-condition of the Vandermonde problem and the roundoff errors.

Figure 6 shows an example when curve fitting by one single polynomial function reaches its limit in simulating curve sections containing sharp curvature changes. In this example, there are two sharp points. For the curve segment containing the left sharp corner where the sample starts and ends, the fitted curve matches the original curve well. At the other sharp point, the fitted value is smooth and cannot reach the sharp corner as shown in Figure 6. The third order polynomial functions were used. For a continuous and smooth curve as shown in Figure 7, although the curve has an intersection point that makes two closed ellipse shapes, the fitted curve can still match the original curve well with the 3rd order polynomial functions estimated using the moving least squares method with C=2.0.

Fitting accuracy and random error

Generally speaking, the order of polynomials used is directly related to curve fitting accuracies as can be seen in Table 1. Because of the limitation of computer precision, when the order was 11 or greater in our experiment, the fitting accuracy decreased dramatically using the least squares. Moving least squares resulted in higher accuracies of curve fitting with lower order polynomials.

Table 2 is a comparison of errors among the original curve, the digitized curve and the fitted curve from

Table 2. A Comparison of Position Deviations

	Position Deviation (cm)
Digitized vs. Original	0.0193
Fitted vs. Original	0.0096
Digitized vs. Fitted	0.0120

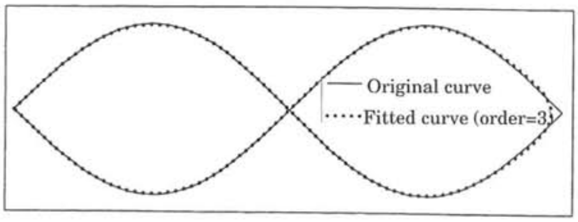


Figure 6. Fitting a curve with internal intersection and abrupt curvature changes. The fitted results were obtained using the 3rd order polynomials estimated with moving least squares.

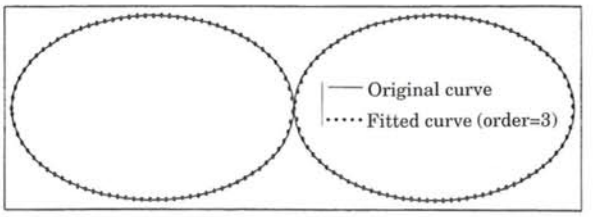


Figure 7. Fitting a curve with internal intersection with no abrupt curvature changes. The fitted version was produced by the 3rd order polynomials estimated from moving least squares.

an experiment. The error of a digitized point was estimated from the minimum distance between the point to the original curve. It can be seen that the standard deviation between the digitized curve and the original curve is 0.193 mm while the standard deviation between the fitted curve and the original curve is 0.096 mm. Thus, the fitted curve has a higher accuracy than the digitized curve (Figure 8). Since the original true line may not be available in practice, the deviation of 0.120 mm between digitized points and the fitted curve may be taken as an approximation of the uncertainty introduced by digitization. Although the approximation tends to be smaller than the true digitizing error, it seems to be a proper measure of curve uncertainty for the application of the epsilon band model because the majority of the true curve will be within a 0.24 mm zone centered at the fitted curve.

Uncertainty modeling in map overlay

Map overlay is an important tool in geographical analysis. Through different operations such as intersection, matching, and merging (Pullar and Beard, 1990), two or more multisource data sets can be combined into one. Because different thematic maps are made by different people at different times using different data sources, the polygon boundaries in different digital maps do not exactly match. Even if the same polygon boundaries are separately digitized, the resultant maps will not exactly coincide due to digitization and other errors. Therefore, the operation of map overlay will generate many small spurious polygons (Goodchild, 1978). Spurious polygons are another source of uncertainty in spatial data bases.

To remove those spurious polygons, three strategies have been used: (1) randomly choose one side and delete the other side; (2) use a straight line to connect the two end points; (3) choose the line that has a higher accuracy or that is from a larger map scale and erase the other (Zhang et al, 1993).

Polynomial curve fitting can be used as the fourth strategy to find a new line as an estimation of the true boundary line based on weighted least squares using all the points of a specific boundary from every layer. The new line has the minimum error if all the layers have the same accuracy. When the relative accuracies are different among layers to be overlaid, weights can be assigned to points in each layer in the weighted least squares estimation. The weight assigned to each layer should be made in

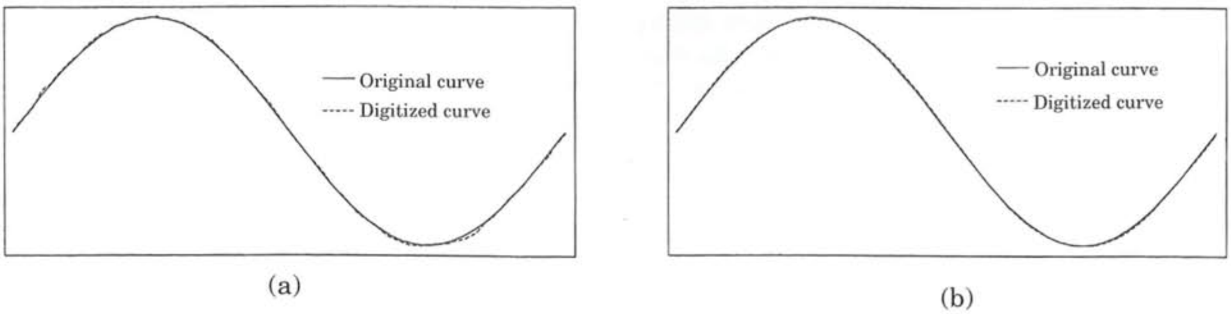
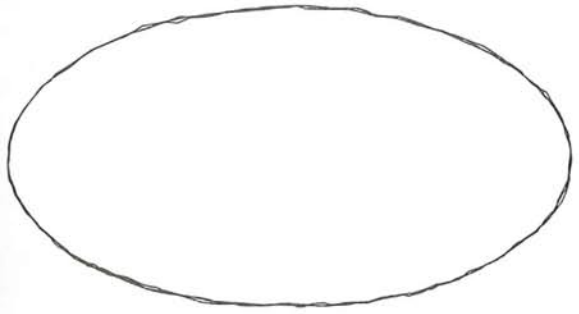


Figure 8. A comparison of the accuracies between the original, a digitized, and a fitted curve. (a) The original curve and the digitized curve. (b) The original curve and the fitted curve.

accordance to the accuracy of the layer, i.e., assign the layer of higher accuracy with a greater weight for the generation of the new boundary. The determination of weights should also be based on the scale of the source maps because maps of smaller scale tend to have less accurate boundary positions.



(b)

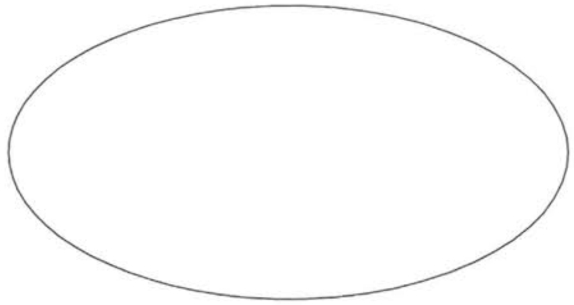
Figure 9. An example of multilayer curve fitting using the polynomial technique. (a) Original ellipse. (b) Three digitized versions. (c) Fitted curve with the 9th order polynomials estimated using weighted least squares technique.

Figure 9 shows some simulated results. Three curves digitized from the same original curve were regarded as three boundary lines each from a different source map. The fitted curve was calculated from the three digitized curves each having the same weight assignment. Table 3 lists the standard deviations between the fitted curve and the true curve and between the digitized points and the true curve.

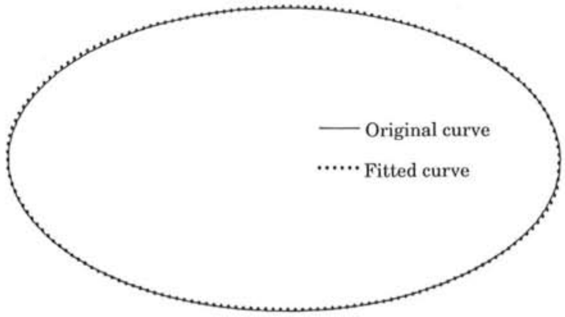
An ellipse curve was produced from a mathematical equation and treated as the true curve (Figure 9a). It was digitized three times and the three digitized versions were overlaid (Figure 9b). Were they displayed at some larger scale, we would see many spurious polygons from the overlaid results along the boundary. Figure 9(c) shows the derived line using the 9th order polynomial functions based on all the points on the three different digitized curves with weighted least squares.

Table 3. A Comparison of the Position Deviations in a Map Overlay Experiment

	Position Deviation (cm)
Digitized vs. Original	0.0167
Fitted vs. Original	0.0078
Digitized vs. Fitted	0.0145



(a)



(c)

VII. SUMMARY AND CONCLUSIONS

In vector-based GIS, linear features are sampled in the form of discrete point series and represented by consecutively joined straight lines. The sampled points contain a large amount of errors and the representation method is not suitable for curve features. Few efforts have been made to improve this situation. The primary objective of this research was to seek appropriate methods to model linear features sampled in spatial databases and to determine uncertainties associated with the samples. We presented some methods based on a decomposition model implemented through parametric polynomial functions determined by least squares and moving least squares techniques to achieve the objective particularly for modeling continuous and smooth curve features.

A discrete point series can be decomposed into three components, the original data set, the systematic error, and the random error. Since the systematic error component may be detected through visualization and calibrated or removed through manual editing, we excluded it from our experiments. If the point series comes from a continuous and smooth curve such as a stream, contour, or a boundary of natural phenomena, through experiments we demonstrated that digitized point series can be represented with polynomial functions. A single polynomial function whose coefficients are determined by the least squares technique or a group of polynomial functions whose coefficients are estimated with the moving least squares technique can be used to model a continuous and smooth curve.

In our experiments, we assumed that the random error of sampled points has a stationary normal distribution. Because a true curve is often unknown in real spatial databases, it is impossible to calculate the sample errors such as errors caused by digitization. We demonstrated the use of least squares and weighting least squares methods for estimating sample errors. The standard deviation between point data sampled along a curve and the fitted curve can be used in an epsilon band model to model uncertainties of the sample data, particularly digitized data.

The order of polynomials required to accurately model a curve is lower for moving least squares than that for the regular least squares. While the moving least squares technique gives higher curve fitting accuracies than regular least squares, it lacks computational efficiency.

These techniques may be used to estimate the uncertainties in map digitization, field survey using GPS units, multilayer map overlay, and to represent curves in spatial databases. Modeling and representing lines with polynomials have potential advantages in saving storage space, curve generalization, curve matching and object recognition.

Selecting suitable models and base functions for linear feature modeling and representation and developing appropriate uncertainty estimation methods for linear features in spatial databases warrant more research attention. Further test of the methods proposed here through experiments with curve features digitized from a map or collected in the field may provide important insights for better spatial data modeling and uncertainty estimation.

REFERENCES

[1] Blakemore, M.. 1984. Generation and error in spatial databases. *Cartographica*. 21:131-139.

[2] Bolstad, P.V., Gesler P., and Lillesand T.M.. 1990. Positional uncertainty in manually digitized map data. *International Journal of Geographical Information Systems*. 4(4):399-412.

[3] Brunsdon, C. and Openshaw, S.. 1993. Simulating the effects of error in GIS. in *Geographical Information Handling - Research and Application*. Edited by Mather, P.M.. John Wiley & Sons: England, pp. 47-61.

[4] Chrisman, N.P.. 1982. A theory of cartographic error and its measurement in digital data bases. *Proceedings, Auto-Caro 5*, pp. 159-168.

[5] Dierckx, Paul. (1993) *Curve and surface fitting with splines* New York : Oxford University Press.

[6] Dutton, G.. 1989. Modeling location uncertainty via hierarchical tessellation. in *Accuracy of Spatial Database*, Edited by M. Goodchild and S. Gopal. Taylor & Francis: New York, pp. 125-140.

[7] Gong, P., and Chen, J.. 1992. Boundary uncertainties in digitized maps I: some possible determination methods. *GIS/LIS'92*, pp. 274-281.

[8] Gong, P., Zheng, X., and Chen, J.. 1995. Boundary uncertainties in digitized maps: an experiment on digitization errors, *Geographic Information Sciences*. 1(2):65-72.

[9] Goodchild, M.F.. 1978. Statistical aspects of the polygon overlay problem. *Harvard Papers on Geographic Information Systems* (ed. G. Dutton) Vol. 6. Addison-Wesley, Reading, Mass.

[10] Goodchild, M.F.. 1990. Modeling error in spatial database, Proceedings, *GIS/LIS'90*, pp. 154-162.

[11] Graupe, D.. 1984. *Time Series Analysis, Identification and Adaptive Filtering*. Robert E. Krieger Publishing Company, Malabar, Florida.

[12] Grogan, T.A., Mitchell, O.R., Kuhl, F.P. and Chhabra, A.K. 1992. A performance comparison for global shape classification between Fourier descriptors and Walsh points methods using simulated data. *Remote Sensing Reviews*, 6(1), pp. 155-182.

[13] Gunther, O. and Wong, E.. 1990. The arc tree: an approximation scheme to represent arbitrary curved shapes, *Computer Vision, Graphics, and Image Processing*, 51:311-337.

[14] Janacek, G. and Swift, L. 1993. *Time Series: Forecasting, Simulation, Applications*.. England: Ellis Horwood Limited.

[15] Keefer, B.J., Smith, J.L. and Gregoire, T.G. 1988. Simulating manual digitizing error with statistical models, *Proceedings, GIS/LIS'88*, pp. 475-483.

[16] Lancaster, Peter. 1986. *Curve and surface fitting : an introduction*. Orlando : Academic Press.

[17] Lin, C.S. and Hwang, C.L.. 1987. New forms of shape invariants from elliptic Fourier descriptors. *Pattern Recognition*, 20(5):535-545.

[18] Maffini, G., Arno, M., and Bitterlich, W. 1989. Observations and comments on the generation and treatment of error in digital GIS data. in *Accuracy of Spatial Database*, Edited by M. Goodchild and S. Gopal. Taylor & Francis: New York, pp. 55-67.

[19] Perkal, J. 1966. On the length of empirical curves. Discussion paper 10, Ann Arbor, Michigan Inter-University Community of Mathematical

Geographers.

[20] Poiker, T.K. 1982. Looking at computer cartography. *Geojournal*. 6(3):241-249.

[21] Pullar, D. and Beard, K. 1990. Specifying and tracking errors from map overlay. Proceedings, *GIS/LIS'90*, pp. 79-87.

[22] Shumway, R.H. 1988. *Applied Statistical Time Series Analysis*. Englewood Cliffs, New Jersey: Prentice Hall.

[23] Wei, W.S. (1990). *Time Series Analysis*. Addison-Wesley Publishing Company: New York.

[24] Zhang, G., Ye, X., and Tan, F. 1993. Multi-tolerance data cleaning technique. *Proceedings of the first Symposium of CPGIS*. pp 224-231.

[25] Zhong, K and Hu, G. 1990. *Digital Signal Processing*. Beijing: Tsinghua University Press.

国家科委副主任徐冠华院士、科学院陈宜瑜副院长、自然科学基金会地学部林海副主任、中国地理信息系统协会李广源秘书长、国家基础地理信息中心陈军副主任等参加了CPGIS的五周年纪念会。