# SIMULATION AND MODEL VALIDATION OF POSITIONAL UNCERTAINTY OF LINE FEATURE ON MANUAL DIGITIZING A MAP

H. S. Wu[a, b], Z. L. Liu[a, *]

[a]Northeast Institute of Geography and Agricultural Ecology Research, CAS, 130012 Changchun - (w_huisheng, liuzhaoli2002)@163.com
[b]Graduate School of the Chinese Academy of Sciences, 100039 Beijing, China - w_huisheng@163.com

**Commission II, WG-II-7**

**KEY WORDS:** Simulation, Digital, Distributed, Model, Quality, Error, Accuracy

**ABSTRACT:**

The line feature is divided into two endpoints and line entity in this paper. The error model of independent point has been studied thoroughly, but study on positional uncertainty of endpoint based on line segment is less. Using the theory of probability and statistics and manual digitization tests, we study the positional uncertainty characteristics of manual digitizing endpoints, and we can get the stable distribution of frequency of digitizing endpoints if repeating about eight hundred times. We find the distribution of positional uncertainty of endpoints and two-dimensional normal distribution is not so conformed. The model of positional uncertainty of line feature has been built based on confidence region or probability statistical method. The error-band model of line segment is verified and established by simulation and statistical method in this paper, which shows bigger at endpoints and smaller at line intermediate and the variance is smallest at middle.Carrying out distribution test of points on the line segment, we obtain the distribution of positional uncertainty of points on the line consistent with that of endpoints, and further verify the feasibility of deducing probability model of line feature based on error propagation theory.

## 1. INTRODUCTION

The uncertainty of line feature on manual digitizing a map, as an important component of spatial data quality in a GIS database, directly restricts the analysis and application of spatial data. Lots of factors contribute to the uncertainty of line feature, and some uncertainties of line feature derive form circumstances of data capture, the main one of which is manual digitizing a map.

It is determined by the law of producing a line that the error distribution of endpoints affects the model of uncertainty of line feature. In the classical theory of measurement, the distribution of point accuracy is seen as the two-dimensional normal distribution, which is the model of error ellipse. The experience distribution of point accuracy was analyzed by experimental methods, which concluded that the error distribution curve of manual digitization showed bell curve which was more intense than standard normal distribution curve (Bolstad et al., 1990). But that was only qualitative conclusion and not good for modeling and error-correction. Some researchers further studied the standardized distribution of point accuracy in theory, and educed that the error distribution was normal distribution (Caspary and Scheuring, 1993). The Kolmogorov test method was used to NL distribution tests that educed the distribution of point accuracy followed NL distribution (Meng et al., 1996), but the form of the distribution function was too complex to model and handle error. Some other researchers concluded that the distribution of positional uncertainty of manual digitization was not necessary normal distribution but P-Norm distribution with p approximately equaling to 1.6 (Liu et al., 1999; Tong et al., 2000; Lan and Yang, 2003). The studies above were all concentrated on the model of independent point, but research on positional uncertainty of endpoints based on line segment was less.

Under the condition that the distribution of point positional uncertainty is normal distribution, the positional uncertainty of line feature has been studied with simulation method and rigorous statistics theory.

The error-band model of line feature is a kind of barred model which circumvolutes true position of spatial entity. The earliest proposed line element error-band theory is ‘ε-band’ model which was a simple symmetric buffer area around a line segment, without considering the relationship of line and points which constitute the line. The theoretical model of ‘ε-band’ was further developed into practical application model (Chrisman, 1982). On practicality aspects, the bandwidth of improved ‘ε-band’ was suggested determined by statistical function of positional errors of line elements accumulated form the fist step to the last step of data collection (Chapman et al., 1997). However, because of the hypothesis that the true position of line must located in the error-band, the error-band model had high sensitivity to the shape of line characters. The bandwidth was estimated by calculating the observed values proportion of lines located in the ‘ε-band’ (Goodchild and Hunter, 1997), but it may be existing two different measuring error values of one line corresponding to the bands with same bandwidth. The model based on confidence region was proposed, which discussed some situations of normal distribution with different parameters of endpoints accuracy, and the confidence region was a kind of band including measuring locations while the true position of line was included in the confidence region with more than presupposed confidence level probability (Shi, 1994, 2005). The uncertainty model of line was inducted as

* Corresponding author.

uncertainty model based on probability distribution according to error propagation theory (Li et al., 1995) On the basis of 'ε-band' model, with probability theory and fuzzy mathematics theory, the 'E-band' model was proposed and developed to 'G-band' model (Caspasy and Scheuring, 1992; Shi and Liu, 1998,2000), and the 'S-band' model was built (Shi, 1998),and the 'C-band' model was proposed (Lan and Tao, 2003). A new type of error-band model, which called 'H-band' model, was developed based on information theory, whose bandwidth was determined by entropy coefficients (Fan and Guo, 2001; Li et al., 2003), but 'H-band' model could not be used when the distributions of position accuracy were different. The studies on the model of positional uncertainty of Line feature above were all concentrated on deduced model in theory while rarely involving test verification.

The error distribution of random lines was described, at the same time, the model improved that the true position of endpoints constructed line feature located in the error region with the mean value of the endpoints position as the core, and line cluster could be simulated by random connecting two endpoints based on the hypothesis that the distribution of point positional uncertainty was normal distribution (Dutto, 1992).

In order to further perfect existing theory of positional uncertainty of line feature, using simulation method; firstly we study the error distribution of manual digitizing endpoints, then study the characters of positional uncertainty of line feature, and verify and develop error model under the condition that the distribution of point accuracy is normal distribution.

## 2. MATERIALS AND METHODS

### 2.1 Test Conditions

The hardware equipment is a desktop computer with 17-inch CRT monitor and 1024 × 768 screen resolution, while the software platforms are Visual Studio and Arc Engine. In order to control the influencing factors, we keep the consistence of the state in obtaining the test data of the same group, including the location relation of the digitizing operators and the computer screen, surrounding environment of accessing the test data.

### 2.2 Test Procedure

Based on ARCVIEW platform, lines in the horizontal and vertical directions are generated as the objects digitized. First, we establish a point layer and set it invisible, then manually digitize endpoints of two lines in turn and repeat it T times as a integrated group test, in the processing of testes, T is set to 100,200,300…and 1000. At length, displaying the manual digitization point layer and zooming in it appropriately, we can obtain the distribution of lattices around endpoints. Carry on the grid processing the lattices produced by manually digitalizing results of ten groups and compute the number of digitizing points in each grid, that is, statistics frequencies of different digitizing points within the different ranges of errors, and then we can obtain forty two-dimensional matrixes. Summarize the row vectors and column vectors of each matrix; we can attain a column sequence-Y direction and a row sequence-X direction. By analyzing the tendencies of arithmetic mean, standard deviation, kurtosis and the skewness according to the change of T, we can seek the value of T in the steady trend of changes, which is the minimum number of samples when the frequency distribution of manual digitization endpoints is nearly stable. Under the condition of the minimum number of samples, we analyze the sequences resulting from digitizing endpoints with distribution test to research the form of the directional error distribution of digitizing endpoints.

Based on the hypothesis that the distribution of point accuracy is normal distribution and two-dimensional coordinate mutually independent, first, we set point A with true coordinates ($X_a$, $Y_a$) and point B with true coordinates ($X_b$, $Y_b$). Second, randomly generate N points with mean values are $X_a$ and $Y_a$ and standard deviation are $\sigma_{Xa}$ and $\sigma_{Ya}$, and randomly generate N points with mean values are $X_b$ and $Y_b$ and standard deviation are $\sigma_{Xb}$ and $\sigma_{Yb}$, (Figure 1). The distributions of the both N points are all two-dimensional normal distribution. Third, line cluster with M lines are generated by randomly connecting two point masses around point A and B. Then, we grid the line cluster, calculate the frequencies of line elements in the direction of plumbing different position of the line cluster, and analyze distribution characters of the frequencies. Finally, we change the value of M to compare the analytical results, and verify a model of positional uncertainty of line feature on manual digitizing a map.

Figure 1. Diagram of one thousand random points with two-dimensional normal distribution; the mean values are (2,2), the standard deviations are both 0.5, and coordinates range form 1 to 3

## 3. RESULTS

### 3.1 The Positional Uncertainty Distribution of Endpoints

Raster physical structure of the CRT displays leads to the results of digitizing endpoints show a regular grid distribution (Figure 2). Extracting the frequency of the digitizing points, and then we can attain lattice data. Summarizing the row vectors of each matrix, we can attain a column sequence ($Y_t$), and summarizing the column vectors to attain a row sequence ($X_t$). Some matrixes have been appropriately adjusted to making row sequence parallel to the direction of the corresponding line segment.

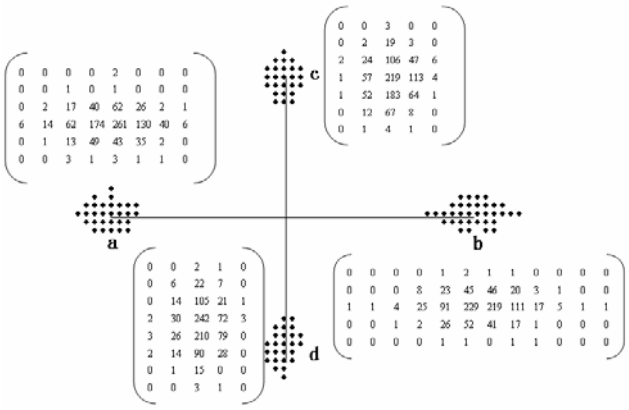


Figure 2. The schematic diagram of manual digitizing endpoints about one thousand times

**3.1.1 The Minimum Samples Needed In Keeping The Distribution Changing Smoothly**: Ten integrated group tests have been done by set T value and made statistical of $Y_t$ and $X_t$ from the digitizing data, endpoint (b) was taken as an example, (Table 1). According to the $X_t$ and $Y_t$ of every endpoint, we calculated the arithmetic mean (m), standard deviation (s), kurtosis (k) and skewness (y), and draw the change tendency map (Figure 3). In view of the frequency of digitizing points, when the positional uncertainty of manual digitizing endpoint is stable, the changing curves of m / T, s / T, k and y of samples should tend to be straight lines with the increase of the number of sample. From the changing maps of the statistical parameters, the minimum number of samples in the process of each test can be easily obtained when the change of tendency is smooth. The result from the analysis of many tests was that the required minimum number of samples was about 800 in studying positional uncertainty of manual digitizing endpoint based on simulation test.

| T | $X_t$ | $Y_t^T$ |
|---|---|---|
| 100 | 0,0,2,4,14,37,30,12,1,0,0,0 | 0,15,57,28,0 |
| 200 | 0,0,2,7,28,59,64,35,5,0,0,0 | 0,44,115,41,0 |
| 300 | 0,0,3,13,39,94,95,47,7,2,0,0 | 0,56,190,54,0 |
| 400 | 0,0,3,16,61,125,121,62,9,3,0,0 | 0,76,252,70,2 |
| 500 | 0,0,4,19,77,154,148,78,14,4,1,1 | 0,92,323,83,2 |
| 600 | 0,0,4,22,93,187,182,89,16,5,1,1 | 1,105,390,102,2 |
| 700 | 0,0,4,24,109,231,208,99,18,5,1,1 | 2,122,454,118,4 |
| 800 | 0,0,4,27,119,267,240,115,21,5,1,1 | 3,129,538,126,4 |
| 900 | 1,1,5,33,134,294,277,127,21,5,1,1 | 4,143,613,136,4 |
| 1000 | 1,1,5,35,142,329,307,150,22,6,1,1 | 5,146,705,140,4 |

Table 1. The $Y_t$ and $X_t$ of manual digitizing endpoint (b) for T times

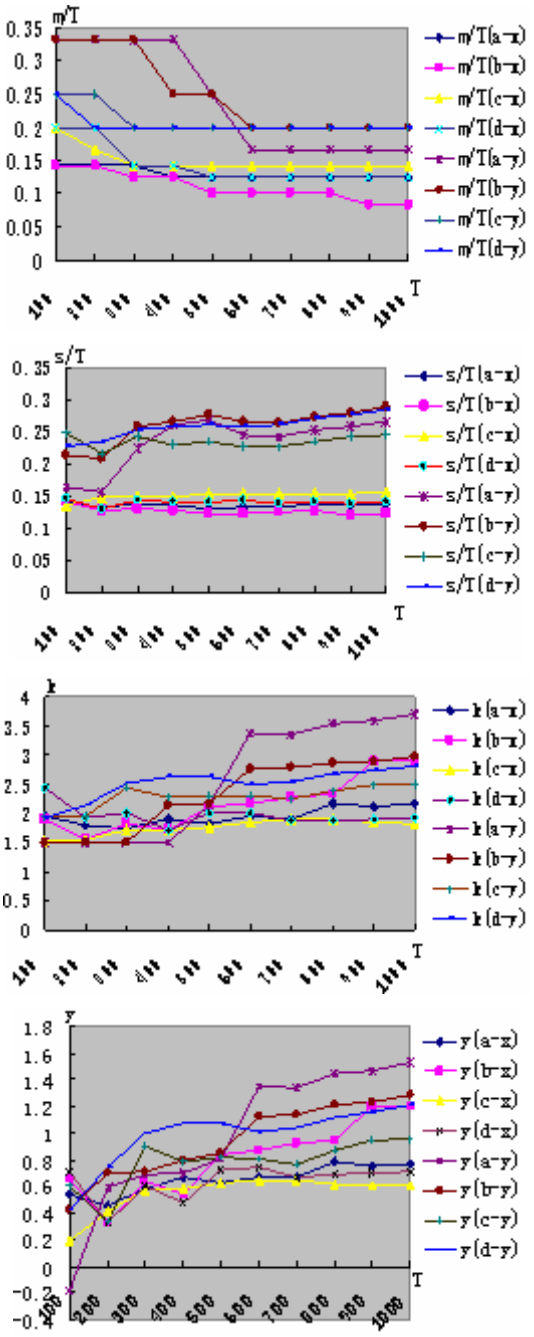


Figure 3. The changing of statistical parameters of sample number on manual digitizing endpoints; (a-x) stands for statistical parameters of $X_t$ at the endpoint (a); (a-y) stands for statistical parameters of $Y_t$ at the endpoint (a)

**3.1.2 The Distribution Test of Positional Uncertainty On Manual Digitizing Endpoints:** According to collecting data of every endpoint by digitizing one thousand times, we can obtained four row sequences ($X_t$) and four column sequences ($Y_t$). All of the sequences were analyzed with Lilliefors test and kstest function test of Kolmogorov-Smirnov based on MATLAB software, and the results of test parameters were described in Table 2. From the test parameters, we could recognize that all sequences were not in accordance with standard normal distribution and some sequences were subjected to normal distribution but others were not with significance level of 0.05. Therefore, the distribution of positional uncertainty of endpoints and two-dimensional normal distribution is not so conformed, and the former has a little skewness.

| | K-S (H) | L (H) | K-S (P) | L (P) |
|---|---|---|---|---|
| Xt-(a) | 1 | 0 | 3.55E-08 | NaN |
| Xt-(b) | 1 | 1 | 1.83E-08 | NaN |
| Xt-(c) | 1 | 0 | 2.98E-07 | NaN |
| Xt-(d) | 1 | 0 | 3.72E-08 | 0.169 |
| Yt-(a) | 1 | 1 | 4.28E-06 | 0.015 |
| Yt-(b) | 1 | 1 | 1.89E-05 | 0.028 |
| Yt-(c) | 1 | 0 | 1.89E-05 | NaN |
| Yt-(d) | 1 | 0 | 1.89E-05 | 0.191 |

Table 2. Statistics of distribution test parameters of $X_t$ and $Y_t$ from manual digitizing endpoints; the significance level is 0.05; K-S stands for Kolmogorov-Smirnov test, L stands for Lilliefors test; (H) stands for test parameter H, (P) stands for test parameter p; if rejecting the hypothesis of normal distribution, H is equal to 1 and P is less than 0.05, otherwise, H is equal to 0 and P is more than 0.05; (a), (b) ,(c) and (d) are all stands for endpoints.

### 3.2 Model Validation of Positional Uncertainty of Line Feature

Although the distribution of positional uncertainty of endpoints is not normal distribution, the distribution model has not been mature. While positional uncertainty of independent point agree with normal distribution which was most previous studies based on , in order to easily modelled and compared with former studies, and based on the density function of two-dimensional normal distribution, we random generated one thousand points coordinates of which the mean value was (200,200) and standard deviation was (50,50) and both the X values and Y values were located between 100 and 300, and random generated another one thousand points coordinates of which the mean value was (800,200) and standard deviation was (50,50) and Y values were located between 100 and 300 while X values were located between 700 and 900. We programmed and random connected two point masses to generate M lines, while M was set as 100, 500, 1000 or 2000. (Figure 4).

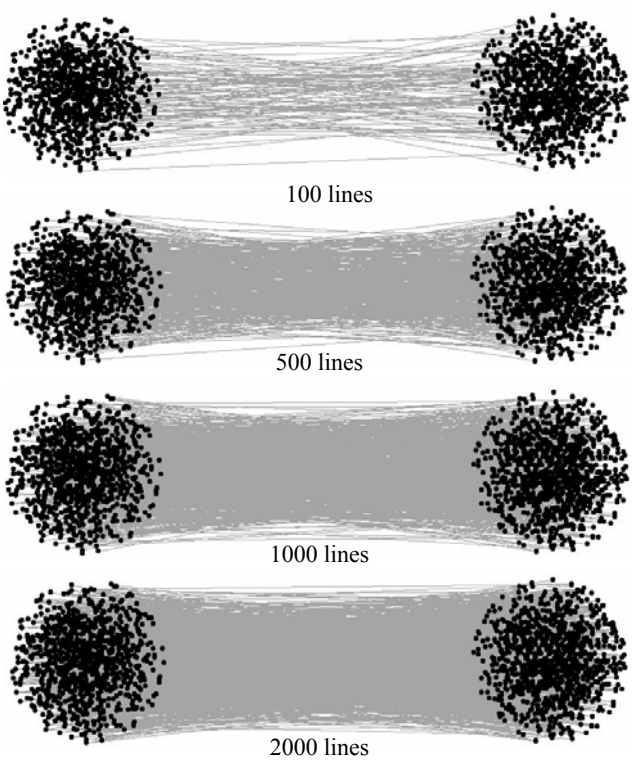


Figure 4. The schematic diagram of generating lines from random connecting two point masses

**3.2.1 The Error-band Model Validation of Positional Uncertainty of Line Feature**: The range of point masses formed a circle surface region, of which the radius was two times of standard deviation in our tests. We selected some point positions which were A(200,200), C1(300,200), C2(400,200), C3(500,200), C4(600,200), C5(700,200) and B(800,200) as characteristic positions of line cluster. The width of line cluster at characteristic positions was measured and recorded in Table 3. In view that the error-band model is more accurate by measuring line cluster with more lines, the line cluster with one thousand line features were selected to measure the width at characteristic positions and built the error-band, (Figure 5). From Figure 4 and Table 3, it could be seen that the width of error-band was bigger at endpoints and smaller at line intermediate, which was as similar as the existing models, but in tests we found the minimum bandwidth was about eighty-five percent of diameter (d) of error circle at the endpoint. The error-band mode of one-dimensional line segment $L_{AB}$ could be expressed as Equations 1. A line is connected by many points, so the $Y_c$ of an arbitrary point $(X_c,Y_c)$ on a line ranges from 0.85*d to 0.95*d, and the value of h is biggest at the endpoints while smallest at the middle.

$$L_{AB} - h/2 \le L_{AB} \le L_{AB} + h/2 \qquad (1)$$

where $h \in [0.85d, 0.95d]$
d =diameter of error circle at the endpoint.

| Position | 100 lines | 500 lines | 1000 lines | 2000 lines |
|---|---|---|---|---|
| A | 138.0 | 182.5 | 178.5 | 189.4 |
| C1 | 163.7 | 176.3 | 171.2 | 178.0 |
| C2 | 155.9 | 162.8 | 170.0 | 173.9 |
| C3 | 152.2 | 148.2 | 169.5 | 172.9 |
| C4 | 152.6 | 144.5 | 170.0 | 178.0 |
| C5 | 157.5 | 167.5 | 175.6 | 180.7 |
| B | 157.8 | 167.9 | 183.2 | 193.2 |

Table 3. The widths of line cluster at special position according to different random lines

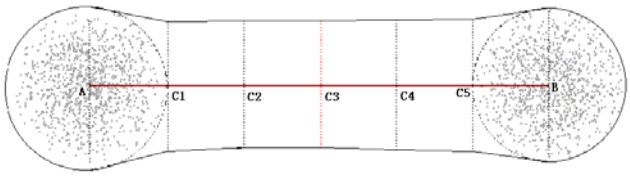


Figure 5. The schematic diagram of error-band of line feature

**3.2.2 The Probability Statistical Model Validation of Positional Uncertainty of Line Feature**: The line cluster was divided into fifty-three equal parts in the horizontal direction and twenty-one equal parts in the vertical direction in the girding process, so grid matrix with twenty-one rows and fifty-three columns were generated. The numbers of line element crossing each grid were recorded, and changed it into frequency by dividing the total number of corresponding column. Seven matrixes with twenty-one rows and one column were generated around the locations of A, C1, C2, C3, C4, C5 and B, and the change law of each column were drawn in Figure 6. It could be seen from Figure 6 that the frequency distributions of symmetrical point positions with the C3 as the centre of symmetry were similar in the vertical direction. The frequency matrixes of special positions of the line cluster including random two thousand lines were analyzed with Lilliefors test, and the line cluster including random one thousand lines, one hundred lines and five hundred lines do the same. The results of Lilliefors test provided that the frequency distribution was normal distribution, so that verified the feasibility of deducing probability model of line feature based on error propagation theory. In view that the change tendency of frequencies can reflect the change tendency of the corresponding probabilities, the frequencies were used approximately instead of the probability , and the curve of frequency changing was smoothed, so a probability statistical model of positional uncertainty of line feature were developed, (Figure 7). The minimum standard deviation of the probability statistical mode was at the middle of a line, while the maximum standard deviation at the endpoint.

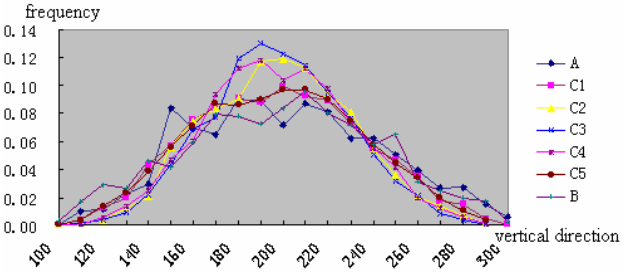


Figure 6. The frequency changing in the direction of plumbing the line at seven spatial positions on line cluster including two thousand lines

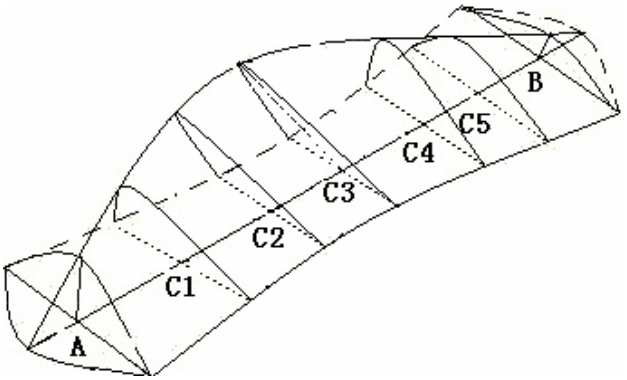


Figure 7. The schematic diagram of probabilistic distribution of positional uncertainty of line feature

## 4. IDISCUSSION

We studied the positional uncertainty of manual digitizing endpoints by manually digitizing testes and statistical probability methods. From the data result, the distribution of positional uncertainty of manual digitizing endpoints, which shows more digitizing points along the direction of a line than the perpendicular direction and more digitizing points in the medial aspect of a line than the lateral, is not in accordance with two-dimensional normal distribution because of the influence of line segment during digitizing endpoints process. Above study was rarely involving in previous researches on the positional uncertainty of line feature of which distribution usually been considered as two-dimension normal distribution. Under the condition that the error distribution of independent point is two-dimensional normal distribution, we studied the positional uncertainty of line feature by statistical simulation. The result shows the bandwidth of line cluster circumvoluting the true position of spatial entity is bigger at endpoints and smaller at line intermediate and the variance is smallest at middle, which is as the same as other research results (Dutto, 1992; Shi, 2005; Dai, 2004). We develop an error-band model of which the shape is similar to the model shape based on confidence region (Shi, 2005), but the bandwidth of our band model is a little bigger, of which the minimum width is about eighty-five percent of diameter of error circle at the endpoint. From girding the line cluster and analyzing the frequencies of line elements at special positions to testing the distribution, we find the distributions of positional uncertainty of points on the line which are all normal distribution are as same as that on endpoints, further verify the feasibility of deducing probability model of an arbitrary point on a line based on error propagation theory. This paper can not be only used to validate of previous theoretical models but also be the theoretical foundation of approximately correcting the systemic error of manual digitization line feature.

## 5. CONCLUSION

Under the condition that manually digitized for more than eight hundred times, positional uncertainties of endpoints distribute with a certain law. The distribution of positional uncertainty of endpoints and two-dimensional normal distribution is not so conformed because of the influence of the line segment, but the distribution model has not been mature, so it needs to be studied thoroughly in the next step.

In order to easily compare with previous research results, based on the hypothesis that the error distribution of independent point is two-dimensional normal distribution, the positional uncertainty of line feature was studied by statistical simulation, and the main character is that the bandwidth of line cluster circumvoluting the true position of spatial entity is bigger at endpoints and smaller at line intermediate and the variance is smallest at middle. The theoretical model of 'ε-band' was verified by testes in this paper. We developed an error-band model based on statistic method. From testing the frequency distribution of line elements at several special positions, the results show the error distributions of points on the line are normal distribution, of which the variances have a relationship with the variances of endpoints and the distance from an endpoint. The feasibility and correctness of deducing probability model of line feature based on error propagation theory are been verified. One of the limitations of this study is that we have not yet analyzed the situation that two dimensional coordinates are correlated and another is that the positional uncertainty model has not been discussed when the positional distributions of endpoints are not normal distribution, both of them need to be studied thoroughly in the next step.

## REFERENCES


Bolstad, P. V., Gesler, P. and Lillesand, T. M., 1990. Positional Uncertainty in Manually Digitized Map Data. *International Journal of Geographical Information System*, 4(4), pp. 399-412.

Caspasy, W. and Scheuring, R., 1992. Error band as measurers of geometrical accuracy. *In: Proceeding of EGIS'92, Utrecht*, pp. 226-233.

Caspasy, W. and Scheuring, R.,1993. Positional Accuracy in Spatial Data bases. *Compute, Environment and Urban Systems*, 17(2), pp. 103-110.

Chapman, M. A, Alesheikh, A. and Karimi, H. A., 1997.Error modeling and management for data in GIS. *In Proceedings of the Coast GIS,97,* Aberdeen, Scotland.

Chrisman, R, A., 1982.Theory of Cartographic Error and Its Measurement in Digital Database[A]. *In: Proceedings of Auto-Carto 5[C], Bethesda, MD: American Congress on Surveying and Mapping*, pp. 158-159.

Dai, H. L., 2004. *Uncertainties Theory and Application in vector GIS*. Science Press, Beijing, pp. 21-94.

Dutto, G., 1992.Handling positional error in spatial databases. *In Proceedings of the 5th International Symposium on Spatial Data Handling*, South Caroline, USA, pp. 460-469.

Fan, A. M. and Guo, D. Z., 2001. The Uncertainty Band Model of Error Entropy. *Acta Geodaetica et Cartographica Sinica*, 30(1) pp.48-53.

Goodchild, M. F. and Hunter, G. J., 1997.A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11, pp. 299-306.

Lan, Y. M. and Tao, B. Z., 2003.Combined Quantification of Line Feature Uncertainties in GIS. *Geomatics and Information Science of Whhan University*, 28(5), pp. 559-561.

Lan, Y. M. and Yang, X. M., 2003. The Distribution Test of Manual Digitization Map Error. *Bulletin of Surveying and Mapping*, 4, pp. 42-44.

Li, D. R., Peng, M. Y. and Zhang, J. Q., 1995.Modeling Positional Uncertainty of Line Primitives in GIS. *Journal of Wuhan Technical University of Surveying and Mapping*, 20(4), pp. 283-288.

Liu, D. J., Shi, W. Z. and Tong, X. H., 1999. *Accuracy Analysis and Quality Control of GIS Spatial Data*. Shanghai Scientific LiteraturePress, Shanghai, pp. 50-120.

Li, D. J., Gong, J. Y., Xie, G. G. and Du, D. S., 2003. The Model of Error Entropy of Area Unit in GIS. *Acta Geodaetica et Cartographica Sinica*, 32, pp.31-35.

Meng, X. L., Liu, D. J. and Zhu, Z. H., 1996.NL Distribution Tests of Map Digitization Errors. *Journal of Tongji University*, 24(5), pp. 525-529.

Shi, W. Z., 1994. *Modeling Positional and Thematic Error in Integration of GIS and Remote Sensing*. ITC Publication, Enschede, pp. 22-147.

Shi, W. Z., 1998.*Theory and Method of Error in Spatial Data*. Science Press, Beijing, pp. 69-125.

Shi, W. Z. and Liu, W. B., 1998.The Stochastic Process Model For Handling Positional Uncertainty Of Line Segments In GIS. *Acta Geodaetica et Cartographica Sinica*, 27(1), pp. 37-44.

Shi, W. Z. and Liu, W. B., 2000.A Stochastic Process—based Model for the Positional Error of Line Segments in GIS. *International Journal of Geographical Information Science*, 14(1), pp. 51-66.

Shi, W. Z., 2005.*Principle of Modeling Uncertainties in Spatial Data and Analysis*. Science Press, Beijing, pp. 84-134.

Tong, X. H., Shi, W. Z. and Liu, D. J., 2000.Error Distribution. Error Tests and Processing for Digitized Data in GIS. *Journal of Wuhan Technical University of Surveying and Mapping*, 25(1), pp.80-84.


## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant No. 40671137.