Boundary Uncertainties in Digitized Maps: An Experiment on Digitization Errors

P. Gong*, X. Zheng* and J. Chen†

*Department of Environmental Science, Policy, and Management,
University of California, Berkeley, CA 94720-3110, USA
†Global Resources and Environmental Assessment Technologies (GREAT)
Albany, CA 94706-1223, USA

Abstract

In this paper, an experiment is described on the study of the relationships between the digitizing error and the curvature. A simulated map was made with 8 circles of different curvatures. Ten operators digitized this map each for 6 times. An average digitizing density was determined for each operator. The original map was used as a reference for digitizing error calculation. Digitizing errors evaluated in this research included positional error, average error band, relative area error and average positional error. In addition, positional error distributions along original circle boundaries were calculated for each operator. The experiment revealed: (1)aera errors and digitizing densities tended to be larger as curvature increases; (2) the mean distributions of positional errors caused by digitization along a circular boundary in most cases did not coincide with the boundary. The results implied that areas of small-sized compact polygons would be underestimated from digitized maps.

摘 要

本文介绍对数字化误差与曲线曲率之间关系进行研究的一个试验。按不同曲率模拟了8个圆，请10位数字化员对模拟圆进行了6次数字化。计算了每个数字化员的数字化点密度。用模拟圆做真值来计算数字化误差。计算了位置误差、平均误差带、相对面积误差和平均位置误差。同时也计算了位置误差与模拟圆比较后所得到的位置误差分布。试验结果表明（1）面积误差和数字化密度随曲率增大而增加；（2）由数字化引起的位置误差平均的分布并不与圆周重合。结果意味着小而形状紧凑的多边形在数字化后面积会被低估。

I. ERRORS AND UNCERTAINTIES IN SPATIAL DATA

Knowledge on errors and uncertainties is critical to the development and application of spatial databases [1-3]. Research on error and uncertainty in the past can be grouped into: (1) classification of errors and uncertainties [2][4-5]; (2) study of uncertainty propagation in geographic models and error modeling [6-11] [27-28]; (3) uncertainty estimation in categorical maps using simulation methods [12-13]; (4) identification of error in GIS data bases [14]; and (5) determination and simulation of digitization errors [15-22].

Error refers to as the deviation from a true value. For example, an error: Δl in length measurement, is defined as L-l where L is the "true" value and l is the measured value [5]. Although it is impossible to obtain absolutely true values from measurements, true values can be used based on mathematical modeling in simulation studies. If true values are not available, uncertainties are surrogates of errors. In this sense, impreciseness, vagueness, and incompleteness of data are all considered as uncertainty

Although different researchers have different definitions on uncertainty [23], we will not attempt to define it in this paper. Researchers have made various classifications of errors and uncertainties [1][4]. Bedard [1] identified four types of uncertainties. These are: 1) conceptual uncertainties, 2) spatial uncertainties, 3) non-spatial uncertainties, and 4) meta-uncertainties. 1) affects the classification of the phenomena. An example is the classification of continuously or naturally varying phenomena into taxonomic or spatial groups. 2) and 3) are related to measurement limitations of the quantitative and qualitative properties of an entity [4]. Measurements are inherently imprecise. They contribute to uncertainties in *spatial* and *non-spatial* attributes of the phenomena. In this context, spatial errors refer to the differences between the "true" locations of certain phenomena and their measured locations in various forms of coordinates. Spatial errors are also called "positional errors" or "geometrical distortions". Examples of spatial error are essentially locational errors of any point and line

features in a map such as contour lines, boundaries of different terrain objects and themes. Non-spatial errors, also called "attribute errors" or "categorical errors", refer to the differences between "true" values of certain phenomena and their measured ones. Examples of non-spatial values are forest density, population, soil salinity, temperature, etc. 4) refers to the degree to which the other types of uncertainties are known. These four types of uncertainties combine together constitute the total amount of uncertainty in a system [1].

Uncertainties associated with a true value is often modeled by a normal distribution [11][17][20][24-25]. Since the true value is not known, the uncertainty of a measured value in relation to its true value is modeled by the same normal distribution (Figure 1). Spatial uncertainty can be treated as a confidence indicator of a particular measured position. Thus, given a specific confidence level, say 90%, according to the distribution one can determine a distance interval such as the epsilon in Figure 1 defining the band into which the true value will fall with a probability of 0.90. The band width may change along a measured line such as a contour line. One can classify the probability levels into ranges and determine their corresponding epsilon intervals and map them accordingly [25].

Although a number of models including the one shown in Figure 1 for spatial uncertainty exist, only a few studies under very specific conditions have been made to determine the actual parameters of these models. Bolstad et al. [15] studied the uncertainty involved in the manual digitization of point data with known coordinates. Dunn et al. [17] investigated the effect of digitizing errors on the area

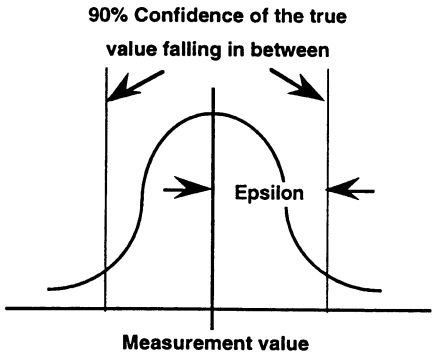

Figure 1. The likelihood of a true value falling into a band centered by a measured value.

estimation of land-use types. In Dunn et al. [17], boundaries in the interpreted land-use map was digitized twice and the polygon segments obtained from different time of digitization were compared to obtain estimates of the uncertainty band. Dutton [24] proposed a two dimensional model for curvilinear map features and made simulations to determine the uncertainty distribution along a line segment. However, Dutton's model is not suitable to model the uncertainties arise from the digitization of a curve. Nonetheless, Dutton has made some interesting suggestions regarding the uncertainty representation. He suggested two ways of uncertainty representation: encoding the uncertainty for each digitized point and use of hierarchical tessellation. Shi [3] expanded the first method to representing the spatial uncertainty of lines.

Many researchers assumed that positional error or uncertainty distribution of a cross-section along a curvilinear map feature was somewhat similar to a normal distribution. For simplicity reasons, an epsilon band was recommended for use to represent uncertainties assuming that the error distribution is symmetrically distributed with the mean coinciding with the true curvilinear feature [26]. However, there have been few studies undertaken to verify the underlying assumption for the use of epsilon band. Two studies proved that the normal distribution assuption was inappropriate. Traylor [22] studied errors involved in the digitization of the boundary of Australia, a generalized scribed line, by a group of operators. A master file was created as reference by carefully digitizing a 25-time enlarged version of the boundary of Australia. Traylor concluded that line-following error can be measured and the associated sequence of errors cannot be considered as a random phenomenon but partially related to the direction of motion of hands. Traylor recommended that line-following performance of an individual could be improved by establishing a training program based on his findings. In another study, Bolstad et al. [15] examined errors of manually digitized point features from topographic maps. They found that the error patterns of the digitized points were not normally distributed. Gong and Chen [18] examined boundary errors or uncertainties from digitized land-use maps produced by several individuals for the same mapping area. It was found that the error or uncertainty in a digitized land-use map is related to boundary curvature and attributes of neighboring polygons sharing the same boundary. They recom-

mended the use of curve fitting and/or blending function to determine boundary uncertainties.

The focus of this study is on the assessment of errors as a function of boundary curvature. Specifically, an experiment has been conducted to examine the relationships between boundary curvatures and area estimation errors, average positional error and digitizing density.

II. EXPERIMENT DESIGN AND METHODS

Map preparation

A set of eight concentric circles and a square frame were generated using a CAD software. The radius of each circle ranged from 0.5 cm to 7.5 cm with an interval of 1 cm. The square frame was 16 cm by 16 cm in dimension. The line width of each circle and the sqaure was 0.08 mm which is a popular line thickness among cartographic products. The circles were printed out on a sheet of standard letter sized paper using a 300 dpi laser printer. Because the dot diameter of the laser printer is 0.085 mm, the actual line width may be slightly thicker than 0.08 mm. The four corners of the square frame were used as control points for the purpose of registration. Root mean square (RMS) registration error is limited to less than 0.03 mm for each operator before digitizing the circle map.

Map digitization

Each circle had a different curvature with the smallest circle having the greatest curvature. The relationship between circle curvature (v) and radius (r) is:

$$v = 360 / 2\pi r$$

These circles were used as the base map for digitizing and as the reference for determining digitizing error. Similar to the experimental design in Bolstad et al. [15], the eight-circle map was mounted on a SummaSketch II digitizing table located in an air-conditioned laboratory. The resolution of the digitizing table is 1000 lines/inch. Each of the ten operators digitized 6 repetitions of this mounted map, three in clockwise direction and three in counter clockwise direction. The operators were asked to keep an operational working speed during the digitization process. This resulted in 60 digitized maps of the original circle map within a time period of 24 hours.

Digitization error determination

Each digitized map was graphically overlaid with the original eight-circle map. The overlay process allows one to visually examine the spatial patterns of digitizing error. For each circle, five types of digitization errors were considered. These were:

- negative area error $\varepsilon_a^- \le 0$, the total digitized area inside a circle;
- positive area error $\varepsilon_a^+ \ge 0$, the total digitized area outside a particular circle;
- relative area error, total area error ε_a divided by the circle area $\varepsilon_a = \varepsilon_a^- + \varepsilon_a^+$;
- average error band $\varepsilon_p = \varepsilon_s / 2\pi r$, where $\varepsilon_s = \left|\varepsilon_a^-\right| + \left|\varepsilon_a^+\right|$;
- positional error ε_i of each digitized point i.

The calculation of ε_a^- and ε_a^+ was carried out by summing up area sections outside and inside the circle, respectively. Positive area section (PAS) and negative area section (NAS) are shown in Figure 2. PAS and NAS are obtained by comparing triangle areas and fan areas in the following forms:

NAS = area ODE + area OEF - fan area ODF
PAS = area OAB + area OBC + area OCD
- fan area OAD

where B, C, and E are digitized points while A, D, and F are intersection points between the digitized lines and the original circle. The average error band can be considered as an error zone centered around the perimeter of a circle. The average error level on each side of a circle can be regarded as half of the band's width. We chose average error band rather than the average positional error calculated from ε_i because we believed that the average positional error would underestimate the level of error when using linked digitized points to represent smooth and round polygons such as circles. It is easy to see that when multi-sided polygons with error-free vertices are used to approximate circles errors still exist along each side of the polygon and reach their maximum at the center of the each side of the polygon.

ε_i is determined for each individual digitized point. The positional error for point "J" (Figure 3) is the length difference between lines OJ and OK (i.e., JK). The length of OK is the circle radius r. In order to compare area errors among circles of different sizes, each of the first three area errors was divided by its circle area. Therefore, the area errors were converted into relative area errors.

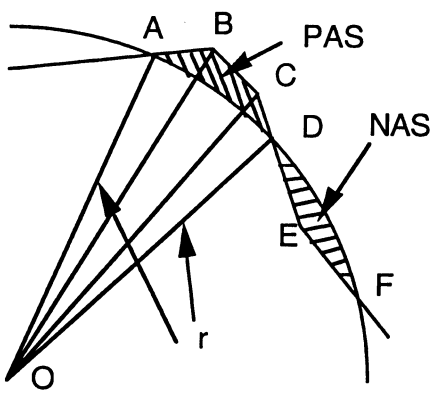

Figure 2. Negative and positive area errors. The polygonal area of ABCD was treated as positive area error while the area of DEF was treated as a negative area error.

Digitizing density

For each digitized version of a circle, the digitizing density was determined through dividing the total number of digitized points by the perimeter of the original circle. The density has been used for evaluating the relationship between digitizing density and curvature.

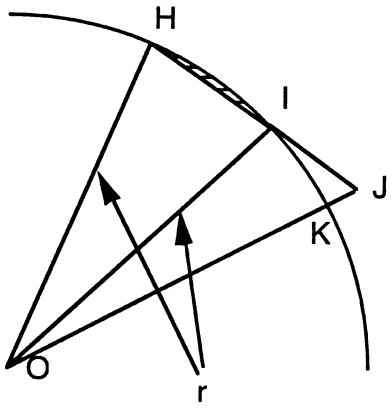

Figure 3. Positional error calculation. Digitized point "J" has a positional error of JK.

Positional error distribution

For each digitized circle, a classification of the magnitude of positional errors can be made. In this study, the error magnitude in the range [-0.155, 0.155] was equally partitioned into 31 pieces with a width of 0.01 cm. The frequencies of positional errors can then be obtained by enumerating the number of points whose errors fall into a specific error range. For each circle, the positional error distribution can be visually examined through a graph generated by plotting the frequencies against the partitioned error ranges

In summary, for each digitized circle three relative area errors, an average error band, a digitizing density and a positional error frequency distribution were determined. For an original circle, these calculations were averaged in each digitization direction for each operator. These averaged results were then examined with respect to circle curvatures.

III. RESULTS AND DISCUSSION

In this section, experimental results on relationships between curvature and average error band, relative area error and digitizing density are presented. According to the curvature and radius relationship, the eight radii from 0.5 cm to 7.5 cm correspond to approximate curvatures 114.65, 38.22, 22.93, 16.38, 12.74, 10.42, 8.82, and 7.64, respectively. If all the curvatures are divided by the smallest curvature, 7.64, the following scaled curvatures are obtained: 15, 5, 3, 2.14, 1.67, 1.36, 1.23 and 1. We will use the scaled curvatures along a logarithmic scale.

For each digitizing direction and each circle, we calculated the means and standard deviaitons for average error band, relative area error and digitizing density. Using t-test, we found most of the differences between the two digitizing directions were insignificant at the 0.90 probability level. Therefore, in the rest of the paper we will mainly present results based on the average of the two digitizing directions.

Curvature versus digitizing density

Digitizing densities are plotted against scaled circle curvatures and a curve is drawn for each of the ten operators at either clockwise or counter clockwise digitizing direction (Figure 4). The highest digitizing density is approximately 2.5 times of the lowest density. The general trend is that the digitizing density increases as the curvature of a circle

increases indicating that digitizing operators tend to take more digitizing points when the curvature is high. However, when the scaled curvatures are small (<3) the digitizing densities do not vary a lot. It can also be seen that there is not much difference in digitizing densities between the two different digitizing directions.

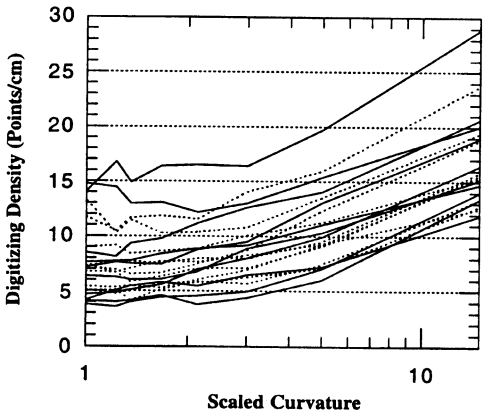

Figure 4. Digitizing density versus curvature. Real and dashed lines represent clockwise and counter clockwise digitizing directions, respectively.

Curvature versus average error band

The relationship between the average error band and the curvature is shown in Figure 5. It can be seen that most of the average error band range from 0.08 to 0.14 mm. This means that the average digitizing error ranges between 0.04 and 0.07 mm on each side of the circle. This is comparable with Rhind's finding that digitizing error can not be much better than 0.075 mm [21]. There is no obvious correlation between the scaled curvature and the average error band for any operator.

Curvature versus relative area error

The effects of curvature on real area estimates can be observed from Figure 6. The general pattern is that as the scaled curvature increases the relative area error decreases. In Figure 6, a positive relative area error indicates the area of the digitized circle is greater than that of the original circle. On the other hand, a negative relative area error means the area of the digitized circle is smaller than that of the original circle. It can be seen from Figure 6 that when the curvature approaches the lower end of the scaled curvature the error magnitude is less than 0.5% and has less variation. In this situation,

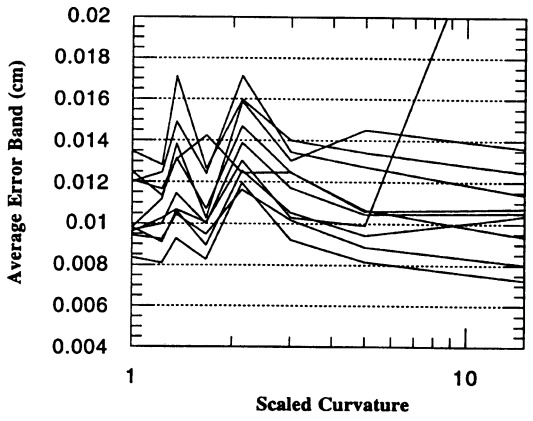

Figure 5. Width of average error band versus scaled curvature.

most of the relative area errors are negative. However, toward the other end of the scaled curvature the magnitude of relative area errors increase at the negative side and vary widely except for the smallest circle. For the smallest circle, the average relative area errors scatter widely apart but within the range between -2% to 2%. The sudden change in relative area error trend with the smallest circle reflects that extra care has been taken by each operator.

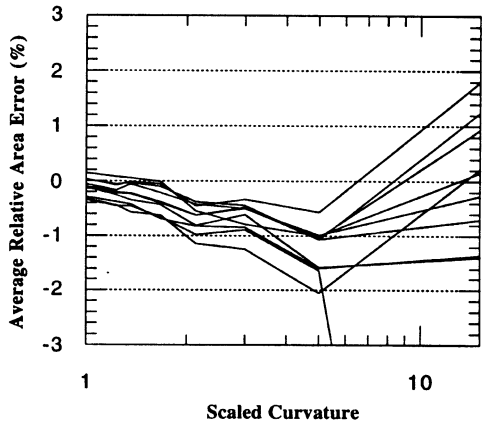

Figure 6. Average relative area error versus scaled curvature.

Positional error distribution

To examine how positional errors are distributed along a circle boundary, positional errors for each

circle and each operator were averaged. The average positional errors were classified into 20 average positional error classes starting with -0.09 cm or poorer and end with 0.09 cm and greater with an interval of 0.01 cm in between. Figures 7 show the average positional error distribution along each circle's cross section for two operators, one with the smallest error range (Figure 7a) and the other with the largest error range (Figure 7b), respectively.

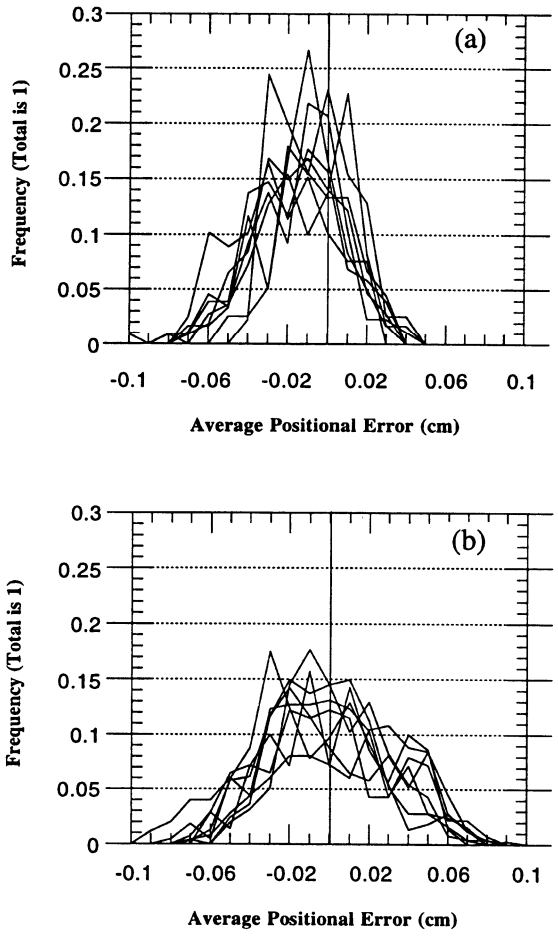

Figure 7. Average positional error distribution along circle's cross section. (a) Results from an operator with a relatively small range of digitization errors; (b) Results from a different operator with a relatively large range of digitizing errors.

It can be seen from Figure 7 that none of the error distribution is normal. In addition, more than 50% of the digitizing errors were made on the negative side, i.e., more points were digitized on the concave side of the circle. Digitizing errors for more than 50% of the digitized points fall within the range between -0.2 to 0.2 mm. This indicates that on each side of the circle the average line following digitizing error were better than 0.2 mm but less accurate than 0.1 mm, the best scenario suggested in [21].

Discussion

When we defined the average error band and expected that half of its width would be greater than the average positional error as directly calculated from the positional digitizing errors along each circle. Our assumption was that most of the digitized points would be very close to the circle. This assumption did not hold in this experiement. The average positional digitizing errors are greater than the widths of average error bands. Although the results are controdictory to our original expectation, we believe that the average error band is an appropriate error measure of the representation error caused by linking digitized points to construct polygons for representation of round-shaped phenomena, particularly circular ones.

Our results indicates that higher digitizing densities neither help reduce positional digitizing errors nor reduce average error bands and relative area errors. These can be partly observed from Figures 4, 5 and 6.

The relatively large negative area errors for circles with large curvatures (Figure 6) indicate that the area of an original circle tends to be reduced after being digitized. This implies that area sizes of small and compact objects presented on a map will be reduced after being digitized into a database. This suggests that in area calculation of digitized polygons it may be necessary to take shape and size information into account to calibrate the area underestimation for those small and compact polygons.

It can be inferred from Figure 7 that (1) a normal distribution assumption is not appropriate in modelling boundary positional errors; and (2) an error model such as an epsilon band centered around a digitized curve is not appropriate for representing positional errors or spatial uncertainties. From the observations in this research, it is expected that positional errors caused by digitization tend to be larger on the concave side of a curve and the magnitude is higher when the curvature is large.

IV. CONCLUSIONS

From this study, it seems that no strong relationships exist between digitizing density, the average error band, relative area error and positional errors evaluated. The results suggest: (1) as the curvature increases, the magnitudes of relative area error and digitizing density also increase; (2) The area of small sized compact polygons may be underestimated from digitized maps, therefore area estimates of small and compact polygons should be corrected by an enlargement factor.

Positional errors caused by digitization along a circular boundary are not normally distributed. For lines of high curvature the mean of positional errors due to digitization tends to fall on the concave side of the curve instead of coincident with the original curve. Therefore, error models such as the epsilon band which centers around the digitized curve will not be appropriate for modeling and representing digitization errors. This type of experiment is useful for testing and training digitizing operators.

REFERENCES

[1] Bedard Y., 1987. Uncertainties in land information systems databases, *Auto- Carto 8,* pp.175-184.

[2] Chrisman, N., 1991. The error component in spatial data. in *Geographical Information Systems*, edited by Maguire, D.J., Goodchild, M.F., and Rhind, D.W., pp. 164-174, Longman:UK.

[3] Shi, W., 1996. Data quality issues in urban geographic information systems, in *Urban Geographic Information Systems: Methods and Applications*, Edited by P. Gong, The Association of Chinese Professionals in Geographic Information Systems (Abroad): Berkeley, California, USA, pp. 69-80. (In Chinese)

[4] Robinson, V. and A. Frank, 1985. About differing kinds of uncertainty in collections of spatial data, *Auto-Carto 7*, pp.440-449.

[5] Maling, D.H., 1989. *Measurements from Maps*, Pergamon Press:New York

[6] Burrough, P.A., 1986. *Principles of Geographical Information Systems for Land Resources Assessment.* Clarendon Press:Oxford, U.K.

[7] Heuvelink, G.B.M., P.A., Burrough, and A. Stein, 1989. Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems.* 3(4):303-322.

[8] Heuvelink, G.B.M., and P.A., Burrough, 1993. Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems.* 7(3):231-246.

[9] Crain, I.K., P. Gong, M.A. Chapman, S. Lam, J. Alai, and M. Hoogstraat, 1993. Implementation considerations for uncertainty management in an ecologically oriented GIS, *Proceedings of GIS'93 Symposium*, Vancouver, B.C., pp. 167-172.

[10] Leenhardt, D., M. Voltz, and M. Bornand, 1994. Propagation of the error of spatial prediction of soil properties in simulating crop evapotranspiration. *European Journal of Soil Science*, 45:303-310.

[11] Shi, W., 1994. *Modelling Positional and Thematic Uncertainties in Integration of Remote Sensing and Geographic Information Systems*, ITC Publication No. 22, Enschede, The Netherlands, 147p.

[12] Fisher, P., 1991. Spatial data sources and data problems. in *Geographical Information Systems*, edited by Maguire, D.J., Goodchild, M.F., and Rhind, D.W., pp. 175-189, Longman: UK.

[13] Goodchild, M.F., G. Sun and S. Yang, 1992. Development and test of an error model for categorical data. *International Journal of Geographical Information Systems*, 6(2):87-104.

[14] Scott, L.M., 1994. Identification of GIS attribute error using exploratory data analysis. *Professional Geographer*, 46(3):378-386.

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

[16] Carstensen, L.W., 1990. Angularity and capture of the cartographic line during digital data entry. *Cartography and Geographic Information Systems*, 17(3):209-224.

[17] Dunn, R., A.R. Harrison, and J.C. White, 1990. Posi tional accuracy and measurement error in digital databases of land use: an empirical study. *International Journal of Geographical Information Systems*, 4(4): 385-398.

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

[19] Lester, M., and N. Chrisman, 1991. Not all slivers are skinny: a comparison of two methods for detecting positional error in categorical maps. *GIS / LIS'91.* Vol. 2, pp.648-658.

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

[21] Rhind, D., 1974. An introduction to the digitising and editing of mapped data. *Automation in Cartography*, The Annual Symposium of the British Cartographic Society, South Hampton, 1973. pp. 50-68.

[22] Traylor, C.T., 1979. *The Evaluation of a Methodology to Measure Manual Digitizing Error in Carto-*

graphic Data Bases. Ph.D. Dissertation, University of Kansas, 117p.

[23] Kruse, R., E. Schwecke, and J. Heinsohn, 1991. *Uncertainty and Vagueness in Knowledge Based Systems, Nemerical Methods*, Springer-Verlag:New York.

[24] Dutton, G., 1992. Handling positional uncertainty in spatial databases. *Proceedings of the fifth international symposium on spatial data handling*. August 3-7, 1992. Charleston, South Carolina, USA, pp.460-469.

[25] Hunter, G.J. and M.F. Goodchild, 1995. Dealing with error in spatial databases: a simple case study. *Photogrammetric Engineering & Remote Sensing*. 61(5):529-537.

[26] Chrisman, N.R., 1982. A theory of cartographic error and its measurement in digital databases, *Auto-Carto 5,* pp. 159-168.

[27] Mark, D.M. and F. Csillag, 1989. The nature of boundaries on "area-class" maps. *Cartographica*. pp. 65-77.

[28] Veregin, H., 1989. Error modelling for the map overlay operation. In *Accuracy of Spatial Databases*, Edited by M. Goodchild and S. Gopal. Taylor and Francis: New York, pp. 3-18.