ASSESSMENT OF DATA ACQUISITION ERROR FOR LINEAR FEATURES

Lysandros TSOULOS, Andriani SKOPELITI
Faculty of Rural and Surveying Engineering - Cartography Laboratory
National Technical University of Athens
H. Polytechniou 9, 157 80 Zographou Campus, Athens, Greece
Tel: +30 +1+772-2730 Fax: +30+1+772-2734
Email: lysandro@central.ntua.gr, askop@central.ntua.gr

Working Group IC-14

KEY WORDS: Accuracy, Data acquisition, Segmentation, Uncertainty, Quality assurance.

ABSTRACT

Cartographic features' accuracy, is the result of a number of cartographic transformations such as data acquisition, generalization etc. These transformations have an overall influence on map reliability, which can be assessed through the development of a model quantifying the relevant error components. The process elaborated in this paper, focuses mainly on the error introduced through manual digitization of linear features. Digitization is a sampling process, resulting to deviation from the true position of the recorded line and the change of the original line shape. It has been acknowledged that positional error due to digitization, depends on the shape of the cartographic line. The successful classification of cartographic lines according to their shape, provides the framework for a realistic approach to the problem. It also leads to quantitative results which will be used for the a posteriori assessment of positional error for each linear feature. In this study an experiment is elaborated to qualify and quantify the influence of line segments complexity on the amount of error introduced. The experiment shows that there is indeed a univocal relationship between those two factors. As soon as this relationship is established, it is shown that in order to develop a valid model for the assessment of data acquisition error, linear features must be a priori segmented in homogeneous parts and subsequently assigned a character. The parametric description of linear features and the segmentation, are based on a methodology developed by the authors.

1 INTRODUCTION

Cartographers and GIS experts distinguish two main sources of error: source map error and operational error. Source map error refers to the accumulated error of the map used as input source. The map producer organization should provide information about source map reliability. When such information is available, it can be utilized in the assessment of the reliability of the resulting data files. Otherwise, the map source is considered error free and uncertainty regarding the data can only be estimated in relation with the map scale. Operational error is a result of data input and data manipulation processes.

Digitization, manual and automatic, is a useful method for data entry. Despite the availability of hardware/software for "automatic" conversion of paper maps into digital form, a considerable part of the digitization of paper maps is still carried out using manual methods. Researchers and practitioners, consider manual digitization as one of the main sources of error in geographic data bases, although it is often largely ignored. Beyond simple checks in the editing process, practical means of handling digitization error are not applied in the relevant procedures. In this study, it is assumed that the positions as they are portrayed on the maps are true, thus this paper focuses on the assessment of error introduced through manual digitization.

2 MANUAL DIGITIZATION ERROR AND LINE COMPLEXITY

Jenks (1981) categorizes human digitization error into psychological and physiological. Error resulting from variation of line thickness and the digitization method followed, can also be considered as an additional factor (Heywood et al., 1998). The true footprint of a line feature lies along its centerline, but it is difficult - if not impossible - for the operator to follow exactly the center of the line. The displacement of the cursor on either side of the centerline is inevitable, leading to the generation of positional errors. This is the line - following aspect of digitization error. In point mode digitization, digitizing is based on the careful selection of sample points which create a faithful representation of a line.

This way, the operator who decides both the number and the location of the sample points, introduces a certain degree of generalization and the corresponding error is the line generalization error. The line following and line generalization aspects of digitization error, lead to the to deviation from the true position (positional error) and possibly to the modification of the true shape of the recorded line.

Error introduced in the data encoding processes is scale dependent. In addition to this, the relationship between data digitization error and line complexity has been identified by a number of researchers (Amrheim et al., 1991; Keefer et al., 1988). Openshaw and Brundson (Openshaw & Brundson, 1993) note: “The key assumption is that error is some function of the geometry of the line as this is the only class of predictor that is generally available”. Abbas et al. (Abbas et al. 1995) state that “the more complex an object is, the bigger gets the probability that it will be erroneous in some parts”. It becomes apparent that error cannot be assumed to be uniform over the map extent or along a single line, which is not homogeneous. As a result, a tool that will permit line segmentation in homogeneous parts will be of crucial importance in the course for the development of an efficient error model.

This study mainly focuses on the assessment of positional accuracy due to manual digitization, in conjunction with the retention of line character. The approach followed is deterministic and the amount of error introduced is related to the shape of cartographic line. An experimental analysis is conducted to identify the relationship between the amount of error and the complexity of the cartographic line/segment.

The analysis utilizes the methodology for partitioning linear features in homogeneous segments based on fractal dimension and the parametric description of the shape of cartographic lines, developed by the authors (Tsoulos and Skopeliti, 1999). The linear features used as experimental data, initially undergo the segmentation and parametric line description procedures. Subsequently, linear features segments are grouped into clusters on the basis of their complexity. These groups, representing varying degrees of features complexity, will be examined in relation with the data acquisition error.

The method described for the assessment of data acquisition error and its relation to line complexity, is consistent with the method for the assessment of line generalization error developed by the authors (Tsoulos and Skopeliti, 2000). This way a framework for the assessment of the quality of linear features in the map making cycle is established, based on measures of positional accuracy and shape retention. This approach utilizes and interrelates three knowledge elements. Provided that linear entities are described by their geometry, structural knowledge is acquired through shape measurement techniques. Structural knowledge describes the line shape and assigns a character to each cartographic feature. The absence of gross error and the minimization of error, is ensured by the contribution of the procedural knowledge. Procedural knowledge specifies the correct methodology and contributes to the assessment of error. As a result, the utilization of geometrical, structural and procedural knowledge, lead to the formation of a model for the assessment of error and the overall quality of the map.

3 METHODOLOGICAL APPROACH

3.1 Segmentation

In order to control the shape variation of the cartographic line, it has to be segmented into homogeneous segments. Segmentation is accomplished through a methodology developed by the authors, which refers to natural features and is based mainly on fractal dimension variation (Tsoulos & Skopeliti, 1999). This methodology is implemented in three steps:

a. Self - similar segments are identified along the linear feature
b. Groups of self - similar segments are formed using cluster analysis
c. A representative segment is selected for each group of segments (the one with fractal dimension value closer to the average value of the cluster).

Four parameters are used for the description of line shape: fractal dimension, average magnitude angularity (Bernhardt, 1992), error variance and the ratio of length and the base line length (Buttenfield, 1991). These parameters are calculated for each line segment and hierarchical cluster analysis is conducted to identify the number of groups that exist in the data. Finally, line segments are classified with non hierarchical analysis.

3.2 Data acquisition quality assessment

Data acquisition quality assessment is based on the positional accuracy of linear features and change of shape. Data collected through the manual digitization process is compared to the original line representing the true position and shape. The original data set will be referred from now on as “nominal”.

Shape modification is examined at the line segment level. Hierarchical cluster analysis is conducted to identify the number of groups that the collected data are classified to. The results of this classification indicate whether complexity differentiation between line segments is preserved. The results of the non - hierarchical classification of collected segments, using the initial clusters centers, are used to identify possible changes of shape through comparison with the shape of the nominal data set.

Positional accuracy is the measure of the deviation from the true position of the recorded line. It is measured by Hausdorff distance, which is considered (Abbas et al. 1995) as a univocal value of the distance between any pairs of lines.

On the contrary, the Euclidean distance (1) defined as the minimum distance between two linear features, cannot be used because it is totally independent of the form and the bandwidth of the linear object (Vauglin, 1997) (Figure1).

$$d_E(A,B) = \inf_{x \in A,\, y \in B} (d(x,y)) \tag{1}$$

The Hausdorff distance between two objects in the finite space is defined by equation (2), where A, B are close sets and d(x, A) the classical Euclidean distance from point x to object B and vice versa.

$$d_H(A,B) = \max\left(\sup_{x \in A} d(x,B), \sup_{x \in B} d(x,A) \right) = \max(d_{AB}, d_{BA}) \tag{2}$$

Hausdorff distance can be measured not only for vertices but for every point on the polyline. Calculation of this distance is based on the transformation of the line into a finite set of points. The accuracy of this method depends mainly on the resolution used.

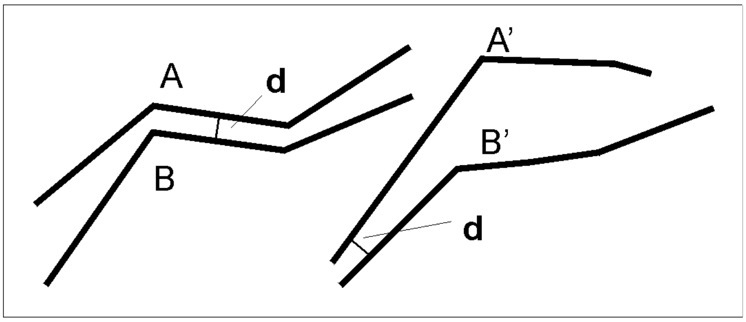

Figure 1. Line segments A , B and A’, B′ have the same horizontal deviation d (Vauglin 1997)

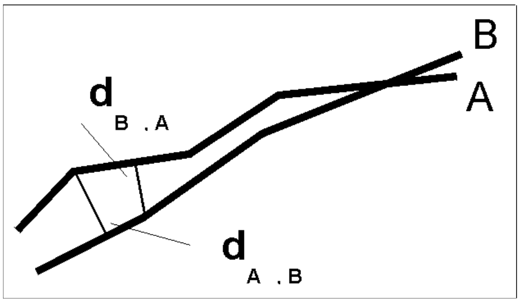

Figure 2. Hausdorff Distance definition

4 CASE STUDY

4.1 Data acquisition

An experiment was conducted to study the results of heads-up digitization, a well known method of data encoding. Heads-up digitization permits the operator to zoom into the line and thus acquire a more detailed encoding of the line footprint. On the other hand, this facility can cause problems, if the operator uses a zoom factor greater than the data resolution implied by the map scale. This leads to the collection of an excessive amount of vertices that overestimate line complexity and do not always result to a more accurate recording.

Heads up digitization was conducted by eighteen [18] different operators in the AutoCAD environment using a color raster image of the source map in the background. Instructions were given to the operators working in the CAD environment to avoid vertices congestion (procedural knowledge). The number of operators involved in the experiment although is not big enough, does not influence the validity of the process. The results will be used as indicators of the operators attitude towards line complexity.

The data set used in the experiment, is the coastline of the Greek island Ithaki, derived from a 1:100 000 scale map. The specific source was selected due to the complicated configuration of the linear features. The coastline is segmented into five (5) segments (Figure 3), which are classified into three (3) groups according to their shape: "smooth" (S), "sinuous" (SIN) and "very sinuous" (VSIN).

4.2 Shape modification

The line segments recorded through heads - up digitization, are classified into three groups utilizing cluster analysis. They have the same synthesis with the clusters of the nominal segments. The same synthesis is preserved even when these line segments are classified using non - hierarchical analysis and the clusters centers of the nominal segments. Thus it becomes evident that noticeable changes of line shape are not present.

4.3 Positional error

The Hausdorff distance measures the positional error of linear segments. From Table 1, it becomes apparent that lines belonging to the group of very sinuous segments exhibit the greatest horizontal deviation. On the other hand, sinuous lines exhibit moderate values and smooth line segments exhibit smaller values. It can be concluded that error in data acquisition measured with the Hausdorff distance follows the pattern of the line classification in three groups based on their complexity. The operators participated in the experiment, handled the three line categories - that were identified through parametric line description and cluster analysis - differently. Taking into account the source map scale, it is estimated that positional error is close to the legibility threshold and no gross errors exist.

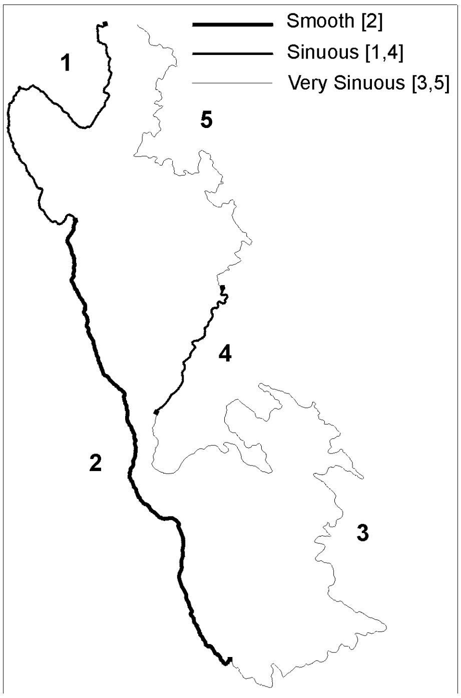

Figure 3. Line segmentation and clustering

Line Code	Haussdorf distance	Group
1	24.51	SIN
2	18.53	S
3	37.73	VSIN
4	23.40	SIN
5	35.97	VSIN

Table 1. Heads–up digitization results

5 CONCLUSIONS

The above mentioned results, show that there is a pattern between positional error due to manual digitization and linear features complexity. Thus in order to develop a valid model, segmentation of lines in homogeneous parts is a prerequisite. Future research on this subject will involve more detailed examination of positional error and the attempt for model development. Research in this field has been done by Vauglin (1997).

The successful classification of cartographic lines according to their shape can lead to the retroactive assessment of positional data collection error for each feature. Information on positional error of linear features due to data acquisition is attached to individual objects and stored in the database as metadata information. Thus the spatial database resulting from the digitization process can be enriched with qualitative description of the data.

This kind of information is indispensable in order to develop a model for the quality assessment of map as a product and the reliability of spatial analysis results using the map as source. On the other hand, the user will be able to decide on the suitability of the data for the intended application/purpose.

REFERENCES

Abbas, I., Grussenmeyer, P., Hottier, P., 1995. Controle de la planimetrie d' une base de donnes vectorielles: une nouvelle methode basee sur la distance de Hausdorff: la methode du controle lineaire. Bul. S.F.T.P. No137 (1995-1), pp.6-11.

Amrhein, C.G., Griffith, D.A., 1991. A statistical model for analyzing error in geographic data in an information system. Discussion Paper No 38. University of Toronto, Toronto, Ontario

Bernhardt, M.C., 1992. Quantitative characterization of cartographic lines for generalization. Report No. 425, Department of Geodetic Science and Surveying, The Ohio State University, Columbus, Ohio.

Buttenfield, B., 1991. A rule for describing line feature geometry. In: Map Generalization, B. Buttenfield and McMaster, R., Eds., Longman Scientific, Halow, Essex, U.K.:, pp.150-171.

Heywood, I., Cornelius, S., Carver, S., 1998. Data quality issues. In: An Introduction to Geographical Information Systems, Longman Pub Group, pp. 178-98.

Jenks, G. F., 1981. Lines, computer and human frailties. Annals of the Association of American Geographers, 71(1), pp.142-7.

Keefer, B.G., Smith, J.L., Gregoire, T.G. ,1988. Simulating manual digitizing error with statistical models. In: Proceedings of GIS/LIS '88, San Antonio, pp.475-483

Openshaw, S., Brundson, C. F., 1993. Simulating the effect of error in GIS. In: Mather, P. (Ed.). Geographic Information Handling. John Wiley and Sons.

Tsoulos, L., Skopeliti, A., 1999. Exploiting parametric line description in the assessment of generalization quality. In: Proc. 19th ICA International Cartographic Conference, 14-21 August, 1999, Ottawa, Canada, vol.1, pp. 1185-1193.

Tsoulos, L., Skopeliti, A., 2000. Developing a Model for Quality Assessment of Linear Features. 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, July 12-14 2000, Amsterdam, Netherlands

Vauglin, F., 1997. Modeles statistique des imprecisions geometriques des objects geographiques lineaires. These de doctorat de l' Univerite de Marne-La-Vallee.