Estimation of digitizing and polygonal approximation errors in the computation of length in vector databases

Jean-François Girres
COGIT Laboratory
Institut Géographique National
Saint-Mandé, France
jean-francois.girres@ign.fr

Patrick Julien
MATIS Laboratory
Institut Géographique National
Saint-Mandé, France
patrick.julien@ign.fr

Abstract— **In vector databases, errors are generally the result of different causes. In this research, we tend to identify the different components of errors, in order to estimate their impact on basic measurements (length, area). This article focuses on two sources of errors impacting length computation in linear vector databases: digitizing error and polygonal approximation. The expression of the impact of these sources of error on length computation is exposed and illustrated by experimentation on the road network of a French topographic database. Results show that the proportion of digitizing error decreases when the total length experimented increases.**

Keywords: errors; length computation; vector datbases; accuracy

I. INTRODUCTION

The description of errors affecting spatial data and their impact was early introduced by Chrisman (1984). Burrough (1986) also proposed to describe the main sources of errors in spatial data. Later, several models have been developed to describe or visualize uncertainty in spatial data (e.g. Hunter and Goodchild, 1996; Heuvelink, 1998; Fisher, 1999). Today, the increased ease of access of geographic data reinforces the users need to receive information on the quality, in order to facilitate interoperability and avoid misuse. In this context, the development of models allowing users to evaluate the quality of them geographic databases becomes relevant.

This research focuses on the assessment of geometric imprecision impact on basic measurements (length, area) in vector databases. To assess this impact, causes of errors have to be understood (section 2). In section 3, two errors impacting length computation are more precisely exposed: digitizing error and polygonal approximation. Thus, experimentation is performed to illustrate their impact on length computation in a topographic database (section 3). To conclude, perspectives are evocated (section 4).

II. SOURCES OF ERRORS IN VECTOR DATABASES

Errors are defined as the deviation from true values. In vector databases, they are generally the result of different causes. In this section, several sources of geometric errors impacting length computation are presented.

A. Projection System

Representations using map projections generate distortions in the representation of the earth surface, and therefore in the computation of lengths. The impact of projection system is well known and can be modelled easily.

B. Geo-Referencing

The geo-referencing of the data support (satellite or aerial imagery, maps...) can provide errors in a vector database after restitution of extracted objects.

C. Terrain Modelling

Computation of horizontal lengths is systematically shorter than using altitudinal information. Even if the altitude is not provided in the vector dataset, it can be extracted from Digital Terrain Model. The impact of the terrain on length computation increases in mountainous areas.

D. Generalization

If a dataset is produced using a map, effects of generalisation also generate errors which impact length computation. As illustrated in Figure 5, several types of errors can be modeled by effects of generalization process: anamorphous, translation, smoothing, and exaggeration

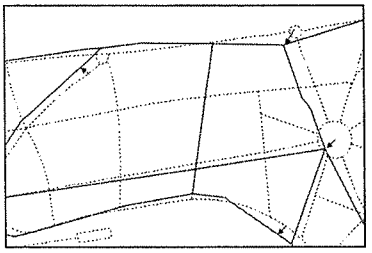

Figure 1. Effects of generalisation on road networks digitizing between BDCARTO® (in plain black) and BDTOPO®

E. Digitizing Error

Digitizing error is generated by the operator during the process of construction of geographic objects (Figure 1). It corresponds to the positional uncertainty around the vertices of a vector object. We consider that these errors are random and independent, and can be modelled by a probability

distribution function. The impact of digitizing error on length computation is developed in the section III.A.

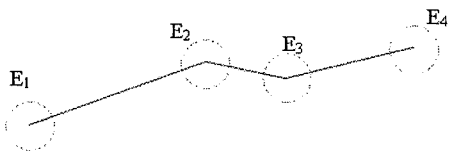

Figure 2. Digitizing uncertainty around each vertex E_n

F. Polygonal Approximation

The polygonal approximation of curves generates a negative and systematic error (Figure 2) on length computation. For a polyline, this error can be estimated by the difference between the polygonal length and the computed length of the curve (section III.B).

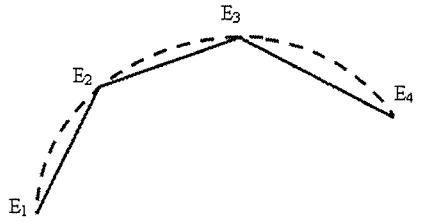

Figure 3. Example of polygonal approximation

G. Other Sources of Errors

Other sources of errors impacting vector databases can be listed, as for instance the precision of the GPS (if the dataset is produced using GPS data), or conversion processes, like vectorization. Siriba (2009) also proposed an approach to estimate positional accuracy of cadastral maps prepared by photogrammetric techniques, using the knowledge of process steps and related errors.

As exposed in this section, several sources of errors can be identified and combined in a linear vector database. However, this paper focuses on two sources of errors: digitizing error and polygonal approximation, in order to study their impact in the computation of length.

III. Expression of Digitizing Error and Polygonal Approximation in Length Computation

Linear objects are represented in a vector database by polygonal lines. The length is *a priori* computed by the length of the polygonal line. Digitizing error can arise because of observation errors by the operator during the capture of linear objects. However, we also have to consider the error due to the polygonal approximation itself when the real object is a curve. The expression of the impact of these errors has been modelled by Julien (2008).

A. Digitizing Error

1) Hypotheses

M1M2...Mn is a polygonal line in the plan, and:

- $m_1 = M_1+\varepsilon_1$, $m_2 = M_2+\varepsilon_2$,..., $m_n = M_n+\varepsilon_n$ are observations on M_i, tainted by error vectors $\varepsilon_i = (\varepsilon_{ix}, \varepsilon_{iy})$.
- norms of errors εi have the same root mean square error (called εq)

We suppose that errors ε_{1x}, ε_{1y}, ε_{2x}, ε_{2y},..., ε_{nx}, ε_{ny} are accidental errors and :

- ε_{ix}, ε_{iy} are normal and centred random variables
- ε_{ix}, ε_{iy} ($1 \le i \le n$) is a family of independent variables
- standard deviation of errors on coordinates x, y are equal : $\sigma(\varepsilon_{ix}) = \sigma(\varepsilon_{iy})$

2) Expression of accidental error e and probability law

We define (1):

$$\frac{M_i M_{i+1}}{\| M_i M_{i+1} \|} = u_i = \begin{pmatrix} \cos\theta_i \\ \sin\theta_i \end{pmatrix} \quad (1)$$

The accidental error (involved by digitizing) on the length of the polygonal line M1M2...Mn is (2):

$$e \cong -u_1 \bullet \varepsilon_1 - \sum_{2 \le i \le n-1} (u_i - u_{i-1}) \bullet \varepsilon_i + u_{n-1} \bullet \varepsilon_r \quad (2)$$

(where • is a dot product)

The error e is a normal centred variable. Its standard deviation σ(e) is (3):

$$\sigma(e) = \sqrt{1 + 2 \sum_{2 \le i \le n-1} \sin^2 \frac{\theta_i - \theta_{i-1}}{2}} * \varepsilon_q \quad (3)$$

where $\theta_i - \theta_{i-1}$ is the angle between consecutive segments $M_{i-1}M_i$ and M_iM_{i+1}.

Properties of the normal law give the following confidence interval (4) with a probability of 99.7 %:

$$-3\sigma(e) \le e \le 3\sigma(e) \quad (4)$$

B. Polygonal Approximation

The error caused by the polygonal approximation of linear curves appears as a systematic error. To assess this error, an estimation of the real length, called "corrected length" needs to be computed.

1) Curvature estimation

M_1, M_2,..., M_n are the successive vertices extracted from a linear curve in the plan.

We consider that the corrected length of the polygonal line $M_{i-1}M_iM_{i+1}$ is the length of the arc of circle passing through the three points M_{i-1}, M_i, and M_{i+1}, as exposed in Figure 4.

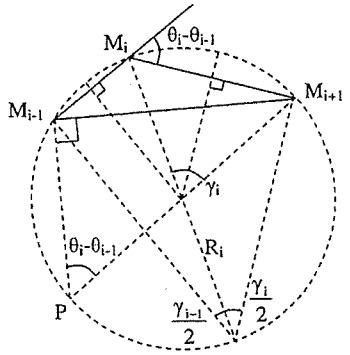

Figure 4. Principle of estimation of the corrected length

By assimilating the curve in the neighbourhood of M_i, to the circle $M_{i-1}M_iM_{i+1}$, we can get the curvature estimation (5):

$$\frac{1}{R_i} \cong 2\frac{\sin(\theta_i - \theta_{i-1})}{\| M_{i-1}M_{i+1} \|} \quad (5)$$

and:

$$\frac{1}{R_i} = \frac{1}{R_2}, \frac{1}{R_n} = \frac{1}{R_{n-1}}$$

2) "Corrected length" computation

We assimilate the arc of curve M_iM_{i+1} to the arc of circle M_iM_{i+1} of radius ρ_i, where (6):

$$\frac{1}{\rho_i} = \frac{1}{2}\left(\frac{1}{R_i} + \frac{1}{R_{i+1}}\right) \quad (6)$$

We suppose that those vertices M_i are close enough each other to admit (7):

$$\frac{1}{R_i} \cong \frac{1}{R_{i+1}} \cong \frac{1}{\rho_i} \text{ or } \frac{1}{\rho_{i-1}} \cong \frac{1}{\rho_i} \quad (7)$$

We can estimate the length of the curve from M_1 to M_n, or "corrected length" by (8):

$$L_c = \sum_{1 \le i \le n-1} \rho_i \gamma_i \quad (8)$$

$$\text{with: } \gamma_i = 2\text{Arcsin}\frac{\| M_iM_{i+1} \|}{2\rho_i}$$

Theoretically, this length is expected to be more accurate than the length of the polygonal line.

We can estimate the error involved by polygonal approximation, using the error b (9):

$$b = \sum_{1 \le i \le n-1} \| M_iM_{i+1} \| - L_c \quad (9)$$

The relative systematic error on the total length of the polygonal line is expressed using the ratio |b|/L, where L is polygonal length of polyline.

IV. EXPERIMENTATION

To illustrate the impact of these two errors on length computation, experimentation is performed using the theme road network of the topographic database BDTOPO© (of metric precision), produced by IGN, the French National Mapping Agency, in the region of Hendaye. This experimentation has been implemented in the GeOxygene library (Bucher et al, 2009).

A. Estimation of errors in Length Measurement

In order to compute the impact of digitizing error on length measurement, the root mean square error ε_q is supposed to be estimated before. Its value varies according to the source and the process used for the capture of the objects.

In this study, the experimentation is only performed on road objects acquired by photogrammetric techniques. Based on the knowledge of data producers, the root mean square error ε_q, which corresponds to the digitizing precision of the operator, can be estimated at 0.5 meters.

Experimentation is performed on a total sample of 1648 linear objects of the BDTOPO®, as illustrated in Figure 5. The total polygonal length of this dataset is 226.74 km.

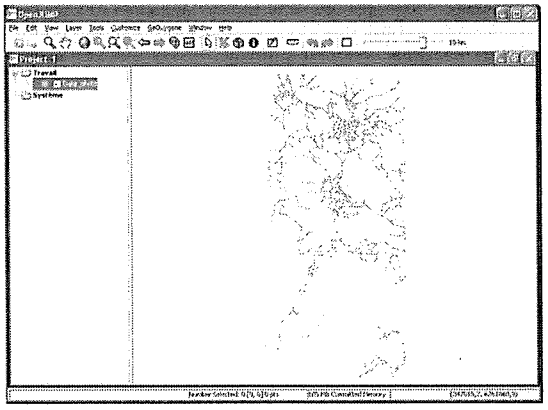

Figure 5. Sample of BDTOPO© road network experimented

Experimentation has been performed using three samples, with a growing number of objects, in order to study the proportion of errors according to the total length.

Impact of digitizing error and polygonal approximation on the total length of the three samples is exposed is Table 1.

TABLE I. COMPUTATION OF DIGITIZING AND POLYGONAL APPROXIMATION ERRORS FOR THE THREE SAMPLES

Objects	L	Lc	3σ(e)	|b|	(3σ(e)+|b|)/L
616	74.88 km	75.03km	0.91 km	0.14 km	1.41 %
1164	145.34 km	145.61 km	1.73 km	0.26 km	1.38 %
1648	226.74 km	227.15 km	2.51 km	0.41 km	1.29 %

Results show that the proportion of errors decreases with the increased number of objects sampled. The impact of polygonal approximation is small and stable ($|b|/L = 0.18\%$), whereas the proportion of digitizing error decreases according to the growing total length experimented.

B. Discussion

The method presented in this paper allows the estimation of digitizing error and polygonal approximation impacts in the computation of length in vector databases. However, different problems are encountered and improvements need to be realized.

The first problem concerns the difficulty to estimate digitizing precision, using the root mean square error ε_q. Knowledge on the capture source and the production process used are necessary to estimate it. The second problem is related to the polygonal approximation, which has been performed on all the objects of the database, but all the road objects are not curve, some of them are sharp objects. Figure 6 illustrates examples of realistic and unrealistic geometric models of curves.

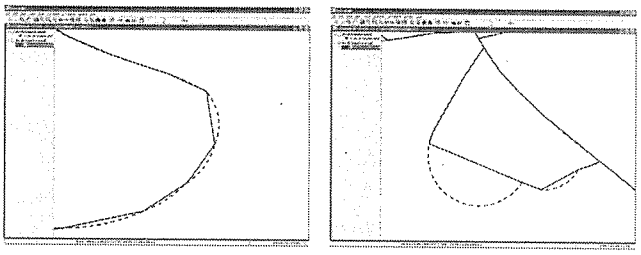

Figure 6. Realistic and unrealsitic computations of curves

An algorithm to detect sharp and curve objects needs to be implemented in order to assess the polygonal approximation impact only on relevant features. In this case, the development of appropriate indicators, such as the sinuosity index proposed by Plazanet et al. (1998), can prove interesting.

V. CONCLUSION AND PERSPECTIVES

This article exposes a formal expression of the impact of digitizing error and polygonal approximation on length computation in vector databases. Experimentation shows that the proportion of the impact of digitizing error decreases when the total length increases. However, the improvement of the model involves the detection of curve and sharp objects, in order to perform assessment of polygonal approximation on appropriate objects. In perspectives, a similar expression of digitizing error and polygonal approximation impacts on area measurement (for polygon objects) will be realized, as far as the modeling of the other sources of errors presented in section 2. These developments are integrated in a general model allowing users to evaluate geometric imprecision impact on vector databases in order to assist decision making.

REFERENCES

Bucher, B., Brasebin, M., Buard, E., Grosso, E. and Mustière, S. (2009). GeOxygene: built on top of the expertness of the French NMA to host and share advanced GI Science research results. *Proceedings of International Opensource Geospatial Research Symposium 2009* (OGRS'09).

Burrough, P. (1986). *Principles of geographical information system for land Ressources assessment*. Oxford: Oxford University Press.

Chrisman, N. (1984). The role of quality information in the long term functioning of a geographic information system. *Cartographica*, 21 (2-3), 79-87.

Fisher, P. (1999). Models of uncertainty in spatial data. In: P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind (Eds) *Geographical information systems: principles, techniques, management and applications* Vol.1. (pp. 191-205) . London: John Wiley and Sons,

Heuvelink, G. (1998). *Error propagation model in environmental modelling with GIS*. London: Taylor and Francis.

Hunter, G. and Goodchild, M. (1996). A new model for handling vector data uncertainty in GIS. *Journal of the Urban and Regional Information Systems Association*. 7 (2), 11-21.

Julien, P. (2008). Note sur les erreurs dans le calcul des longueurs et des aires dans une base de données en mode vectoriel, unpublished.

Plazanet C., Bigolin, N., Ruas, A. (1998). Experiments with learning techniques for spatial model enrichment and line generalization. *GeoInformatica*. 2 (3), 315-333.

Siriba D. (2009). Positionnal accuracy of a cadastral dataset based on the knowledge of the process steps used. *Proceedings of the 12th International Conference on Geographic Information Science (AGILE'09).*