ESTIMATION OF IMPRECISION IN LENGTH AND AREA COMPUTATION IN VECTOR DATABASES INCLUDING PRODUCTION PROCESSES DESCRIPTION

JF. Girres [a], A. Ruas [a]

[a] COGIT Laboratory, Institut Géographique National,73 Avenue de Paris, 94165 Saint-Mandé, FRANCE
(jean-francois.girres, anne.ruas)@ign.fr

Commission VI, WG VI/4

KEY WORDS: Spatial data quality, Geometric imprecision, Production processes, Measurements, Data usage

ABSTRACT:

This paper presents a research on the estimation of the impact of geometric imprecision on basic measurements (length, area) in vector databases, in order to generate relevant information for decision making. The goal consists in the elaboration of a model allowing a non-expert user to evaluate the geometric imprecision of its dataset, using data analysis as well as description of production processes (such as digitising or generalisation). We suppose that these processes induce variable contribution to errors in a dataset, and are exposed to spatial heterogeneity according to the geographical context. This model lays on a knowledge base, based on measurements, contextual indicators and additional information on the dataset production. In order to evaluate a dataset's geometric imprecision impact without any reference, decision rules are under development, using the knowledge base coupled with hypothesis on the influence of the geographical context on production processes. Experimentation on a road network illustrates the respective impact of production processes in the final length measurement error. Possibilities to communicate this impact following a particular usage are also evocated.

1. INTRODUCTION

Since the three past decades, a significant number of research has been conducted on spatial data quality: the description of causes and consequences of errors in spatial databases (Chrisman, 1984; Burrough, 1986), the development of models to describe and visualise error and uncertainty (Goodchild et al., 1992; Hunter and Goodchild, 1996; Fisher, 1999), the development of error propagation models (Heuvelink, 1998) and applications to communicate the impact of spatial data quality for decision making (Devillers, 2004; Ying He, 2008).

In the same period, a global evolution occurred in the production and usages of geographic information: democratization of GIS tools, development of the GIS community in a large variety of users, and more recently the apparition of Volunteered Geographic Information (Goodchild, 2007) allowing to transform anybody in a sensor and a distributor of geographic information. In this context, issues related to spatial data quality and its communication to the final user become relevant.

The COGIT laboratory has been involved in researches on spatial data quality since years (Vauglin, 1997; Bonin, 2002; Olteanu, 2008) and looks for the development of methods and models allowing users to estimate the quality of spatial databases, and its impact on basic measurements. Since 2009, a PhD is conducted in the COGIT Lab, on the conception of a model to evaluate geometric imprecision in vector databases, in order to communicate its impact for decision making.

This paper proposes to expose the approach chosen to elaborate this model, focusing preliminarily on the context and the objectives of this research. A description of the approach to build the model is presented afterwards, illustrated by a practical example, before concluding.

2. CONTEXT AND OBJECTIVES OF THE STUDY

2.1 Spatial data quality

Concepts of spatial data quality

ISO defines *Quality* as the "totality of characteristics of a product that bear on its ability to satisfy stated and implied needs" (Oort, 2005). This definition of spatial data quality includes in fact two sub concepts (Devillers and Jeansoulin, 2006), as presented in Figure 1: *internal quality*, which can be described as the "ability to satisfy specifications" defined by the database producer, and *external quality*, also known as "fitness for use", which corresponds to the needs of the database users.

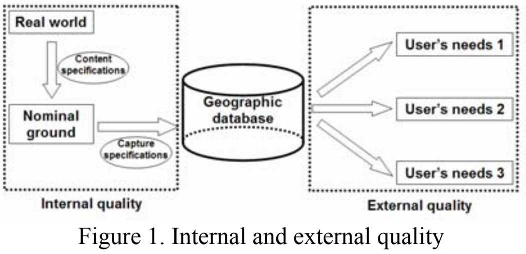

Figure 1. Internal and external quality

Specifications of geographic databases usually satisfy a part of both producer and users concerns. Unfortunately, it appears impossible to integrate all requirements involved by specific usages in the specifications, because they are endless.

Indeed, it is possible to establish a hierarchy of main usages (positioning, or length and area computation) because both concepts of internal and external quality are not completely disconnected. Also, tools and methods have to be provided to the final user in order to communicate the risk involved for a particular usage and to avoid misuses.

Elements of spatial data quality

To evaluate spatial data quality, different characteristics, called “elements” are distinguished. The ISO norm differentiates the following elements (Kresse and Fadaie, 2003): *geometric accuracy, attribute accuracy, completeness, logical consistency, semantic accuracy, lineage and temporal accuracy*.
(Oort, 2005) identified eleven elements in five influent sources publicised since two decades, integrating also *usage*, *variation in quality*, *meta-quality* and *resolution*.

Imprecision, Inaccuracy and Error

Our research concerns the element “geometric accuracy”, but focuses in particular on *geometric imprecision* and its impact on measurements.
Geometric imprecision is defined as the limitation on the granularity or resolution at which the observation is made, or the information is represented (Worboys, 1998a). It has to be differenced to *inaccuracy and error*, defined as the deviation from true values (Worboys, 1998b). For instance, we can estimate the geometric precision in positioning using the Root Mean Square Error (RMS Error).

2.2 Impact of geometric imprecision on measurements

We can consider that the impact of geometric imprecision on length computation can be illustrated by the formula bellow (1):

$$L_{comp} = L_{ref} + \Delta L \qquad (1)$$

where L_{comp} is the compared dataset’s length
L_{ref} is the reference dataset’s length
ΔL is the length variation
L_{ref} is more accurate than L_{comp}

As evocated before, we admit that the use of the RMS Error is well adapted to evaluate geometric imprecision in term of positioning, and looks enough to fit this use. But is it suitable to estimate the impact of geometric imprecision on measurements? To answer this question, a first experimentation using an error-simulation model, following a random law and parameterised with the RMS Error of the dataset, has been performed.

A sample of BDCARTO® road network dataset (RMS Error = 20 m, from specifications) is compared to a simulated BD TOPO® road network dataset (using the RMS Error of the BDCARTO). As presented in the Figure 2, the use of this error-simulation model does not represent faithfully the reality of the exposed example of a road network.

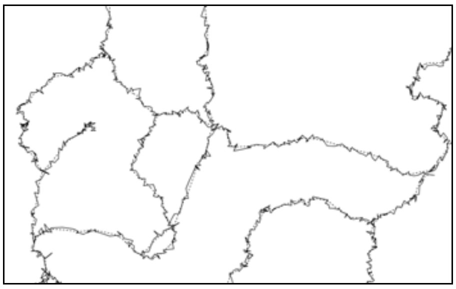

Figure 2. Unrealistic error-simulation on a road network (in plain black) following a random law

Moreover, comparisons of lengths show that this method is not adapted at all to measure the length: The total length of BDCARTO® road network is 67.1 km, compared to the 129.2 km of the simulated BDTOPO® road network, which is totally unrealistic (initial length of BDTOPO® is 65.9 km). This example shows that RMS Error is not suitable to evaluate the impact of geometric imprecision on length measurement.

The problem is quite complex to resolve and we suppose that the comprehension of causes of errors has to be introduced, using description of the different processes potentially considered as sources of errors. The formula below presents the contribution of production processes in the final length deviation (2):

$$\Delta \mathrm{L} = \sum_{i=1}^{n} \Delta P \qquad (2)$$

where ΔP is the variation caused by a production process

We also suppose that local variations of geographical context can affect the contribution of each process generating the final “aggregated error”, because observations show that ΔL presents spatial heterogeneity in the entire dataset. If adding these values is pessimistic, at least it gives a boundary value.

In this context, processes potentially considered as sources of errors have to be understood and modelled, according to them sensitivity to the geographical context.

2.3 Causes of errors in basic measurements

Main sources of errors, in vector databases have been introduced by (Burrough, 1986). We focus here on five sources of geometric errors impacting the computation of length or area: digitizing errors, polygonal approximation, projection system and georeferencing, terrain modelling and generalisation.

Digitizing errors

Digitizing error is generated by the operator during the process of construction of geographic objects (Figure 3). It corresponds to the position uncertainty of each vertex of a vector object.

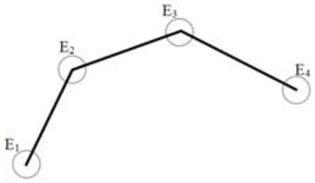

Figure 3. Digitizing error

Digitizing error is a random and independent error, modelled statistically by a probability distribution function (Gaussian Law). Its impact on the length computation of a polygonal line $E_1E_2\ldots E_n$ is modelled by the standard deviation bellow (3).

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

where $\theta_i - \theta_{i-1}$ is the angle between consecutive vectors $E_{i-1}E_i$ and E_iE_{i+1} and ε_q is the digitizing precision

Properties of the Gaussian law give a confidence interval of 99.73% between -3σ(e) and 3σ(e).

Polygonal approximation

The polygonal approximation of curves generates a negative and systematic error (Figure 4) on lengths and areas. For a polyline, this error can be estimated by the difference between the polygonal length and the computed length of the curve.

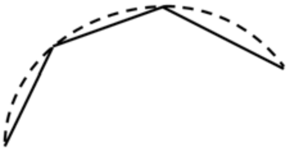

Figure 4. Polygonal approximation of curves

Projection system and Georeferencing

Representations using map projections generate distortions in the representation of the earth surface, and therefore in the computation of lengths. The scale error, defined as the difference between the distance on the map (particular scale) and the distance on the ellipsoid (Principal scale), is used to evaluate the impact of projection on length computation.

In the same time, the georeferencing of the data support (satellite or aerial imagery, maps…) can provide a systematic error in the dataset, after digitising. Parameters like translation, rotation and homothetic transformation have to be estimated.

Terrain modelling

Computation of lengths and areas in two dimensions are systematically smaller than using altitudinal information. Even if the altitude is not provided in the dataset, it can be extracted from Digital Terrain Model. Because the impact of the terrain can be important (especially in mountainous areas), differences have to be estimated to inform the final user.

Generalisation

If a dataset is produced using a map, effects of generalisation also generate errors which impact length and area computations. For instance, road may be translated, sinuous road are smoothed, some bends are removed, or houses may be enlarged and translated to facilitate visualisation.

As illustrated in Figure 5, several types of errors can be modelled by effects of generalisation process: anamorphous, translation, smoothing, and exaggeration.

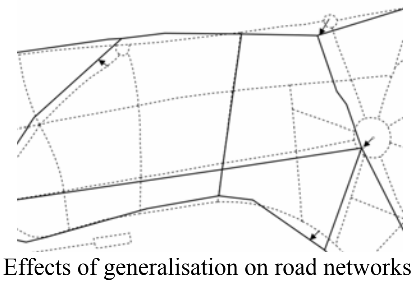

Figure 5. Effects of generalisation on road networks digitizing between BDCARTO® (in plain black) and BDTOPO®

Information on the potential impact of generalisation needs to be provided, by automatic detection, or using user's knowledge.

3. APPROACH

This ongoing research has multiple objectives. The first one is to understand the contribution of each production process in the final aggregated error, according to a particular geographical context. The second one consists in combining appropriate indicators to model geometric imprecision's impact on measurements. Thus, we intend to build a system, based on either reference datasets or hypothesis, related to knowledge on production process and on data.
This system supposes the construction of a knowledge base and decision rules, using preliminarily comparison of datasets.

3.1 Construction of knowledge base and decision rules

As exposed in Figure 6, in a first configuration, comparisons between databases DB1 and DB2 are performed in order to estimate deviations Δ in term of position, length and area. The computed rule base calculates uncertainty (on length or surface) based on measurements on data (data and context) and process by means of machine learning.

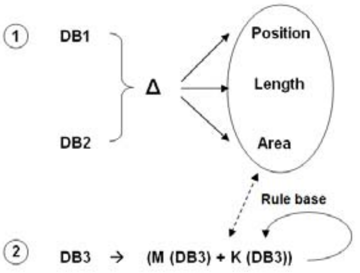

Figure 6. Comparisons between datasets to build the rule base

Then, in a second configuration, for a new dataset DB3 (with no reference dataset), estimation of deviations are computed by means of the rule base, the knowledge on the process (K(DB3)) and measures on the data (M(DB3)). Rule base computation R for a database DB3 is defined as the function (4) below:

$$R\ (DB3) = f\ (\Delta,\ M\ (DB3),\ K\ (DB3)) \qquad (4)$$

where *Δ* are measurements computed using comparisons
M (DB3) is the estimation of deviations
K (DB3) is knowledge on data

Rule base computation supposes to formulate a set of hypotheses. For instance, we can suppose that effects of generalisation are stronger in urban area, effects of terrain are stronger in mountains… Hypothesis formulation involves knowledge on both production process and data. This supposes to collect information provided by the user, and computed using appropriates measurements and contextual indicators (developed in Section 4) integrated in a model (Section 5). Experimentation on a mountainous road illustrates the impact of production processes on length measurements in Section 6.

3.2 Knowledge on production process

Prior the computation of indicators, the user should provide a set of normalised information about the datasets. Most of them are contained in the metadata, but among them, information like the processes used to create the dataset (generalisation or not) or the scales of production and usage have to be provided. This additional information also deals with the confidence on the data sources and processes, and the possible usage.

4. INDICATORS AND MEASUREMENTS

This part presents measurements performed using datasets comparison, but also contextual and shape indicators used to compute the rule base.

4.1 Measurements based on comparisons

To compute measurements, a preliminary phase of data matching of homologous objects is performed. This is realised automatically using algorithms developed by (Mustière and Devogèle, 2008) for linear objects, and (Bel Hadj Ali, 2001) for polygonal objects. Both of them are implemented in the GeOxygene library (Bucher et al., 2009).

Each type of primitives supposes the computation of adapted measurements. For points, indicators of precision (defined as the “fluctuations of a data series around its mean”) and accuracy (defined as the “fluctuations of a data series around the nominal value”) are computed using respectively standard deviation and RMS Error, for X and Y coordinates and deviations (using Euclidian distance). These indicators allow to evaluate the potential bias of the dataset, possible impact of the georeferencing. For polylines, curvilinear abscissa difference, Hausdorff distance and Average distance (Figure 7a) are performed. For polygons, Area difference, Hausdorff distance and Surface distance (Figure 7b) are computed.

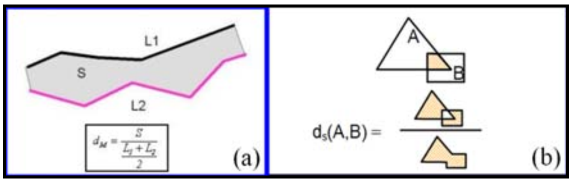

Figure 7. Computation of average (a) and surface distance (b)

4.2 Shape indicators

Shape indicators are computed for linear and polygonal objects. They provide important information to compute rule base. For polylines, indicators of granularity (smallest segment’s length, average segment’s length) and Sinuosity index (Plazanet et al., 1998) are for instance computed. For polygons, indicators of concavity and compactness are also provided. These measurements don’t represent an exhaustive list. Further indicators will be integrated to complete the knowledge on data, as far as contextual indicators.

4.3 Contextual indicators

As exposed in hypotheses, we consider that the weight of the different processes potentially considered as sources of errors, is exposed to variations according to the geographical context. Contextual indicators are produced to characterise the geographical configuration of the dataset and its internal heterogeneity (terrain, density of objects...) in order to determine large areas, like mountains, urban or rural areas. These indicators are produced using the dataset itself, but also external datasets (DTM, networks…). Indicators of neighbourhood are computed in order to evaluate the potential effects of generalisation. For each object belonging to a dataset to evaluate, its neighbourhood population has to be determined. We consider that if this population is important, the object is exposed to effects of generalisation. Distances between objects can provide useful information to determine the scale of generalisation and its potential impact.

5. EVALUATION MODEL

To estimate the impact of geometric imprecision on classical measurements in vector databases, we propose an evaluation model (under development) based on three steps:

- Step 1: Evaluation of a dataset using rule base
- Step 2: Communication of geometric imprecision impact on measurements
- Step 3 : Rule base reinforcement

5.1.1 Evaluation of a dataset using rule base

Two configurations can arise for a user involved in the evaluation of its dataset (Figure 8):

- the user has reference dataset
- the user has no or few reference datasets

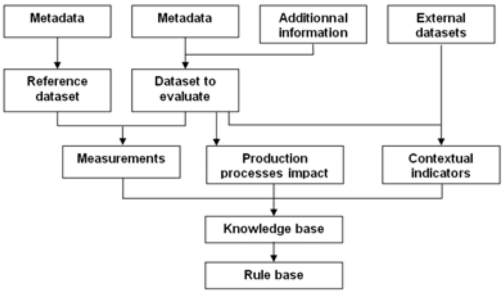

Figure 8. Creation of knowledge base and decision rules

If the user has reference dataset, he can perform comparisons using measurements presented in Section 2, combined with knowledge on production processes and on data. If the user has no, or few reference datasets (like a DTM), geometric imprecision impact on measurements is auto-estimated using decision rules. The rule base is elaborated using previous comparisons integrating knowledge on production processes and on data. Section 6 presents an example of comparison, estimating the respective impact of each production process on length measurement in a mountainous area. Samples of comparisons in different geographical contexts will be performed to elaborate the knowledge base.

Dataset evaluation provides a raw result, not really understandable for the user. In order to fit the use, and communicate clearly the impact of geometric imprecision on measurements, usages have to be taken into account.

5.2 Communication of geometric imprecision impact on basic measurements

Communication of the geometric imprecision impact to the final user involves the introduction of profiles and levels of usage, in order to adapt results to particular usage contexts. Various profiles of users have to be determined, in order to adapt the evaluation in a comprehensive talk. In the same way, thematic example will be used to adapt this communication to particular usage. The goal is to propose as possible to furnish sensitive information to the final user.

5.3 Rule base reinforcement

The last step deals with the reinforcement of the rule base, using validation, or not, of prior evaluations. This revision will be performed manually at the beginning, but we plan to provide a system able to modify rule base according to validated results.

6. EXPERIMENTATION

First experiments are realised to illustrate production processes impact on length computation in linear vector databases, in order to create the rule base. The example focuses on a road network extraction in the mountainous region of Grenoble (France) in two databases: The BDTOPO® and BDCARTO®, produced by IGN, the French National Mapping Agency.

6.1 Presentation of the datasets

The IGN BDTOPO® is a topographic database, of metric positional precision, captured using photogrammetric restitution and ground surveys. The IGN BDCARTO® is a cartographic database, captured using 1:50000 IGN maps and SPOT satellite imagery. Its average positional precision is around 20 meters.
The experimentation focuses on an extraction of road network, the D112 (Figure 9), modelled by polygonal lines in both databases. The projection system used is the RGF Lambert93.

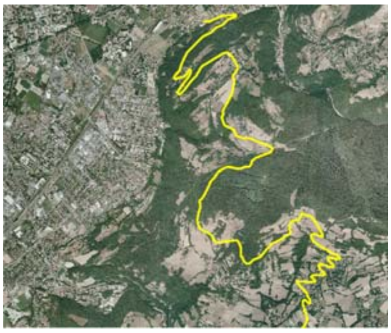

Figure 9. Localization of the road D112, in Grenoble's suburb

Using a classical GIS measurement tool, lengths of the D112 are 12.86 km for BDTOPO® and 12.62 km for BDCARTO®.

6.2 Components of length computation error

The different components of error are exposed hereafter, in order to model them impact in term of length measurement.

Impact of the projection system

Prior to the evaluation of causes of errors (section 2.3), the impact of projection system has to be taken into account.
In the example, the mean scale factor of the road is -0.67 m/km. In consequence, the total length of the road is underestimated of 8.6 meters for BDTOPO® and 8.4 meters for BDCARTO®.

Impact of digitizing error

To model the impact of digitizing error in the measurement of length, the digitizing precision ε_q is defined by the rate between sensibility of the capture (0,1 mm) and the digitizing scale (1:10000 for BDTOPO® and 1:50000 for BDCARTO®). The impact of digitizing error on the total length of the road is expressed by the standard deviation σ(e) in the Table 1.

Dataset	Length	ε_q	σ(e)	3σ(e)
BDTOPO	12,86 km	1 m.	4,9 m.	+/-14,7 m.
BDCARTO	12,62 km	5 m.	30,8 m.	+/-92,4 m

Table 1. Estimation of the impact of digitizing error

Impact of polygonal approximation

Considered as a curve object, the polygonal approximation involves a negative error on the road D112's length. The error b expresses the error in length computation.

Dataset	Length	Corrected Length	Error b
BDTOPO	12,86 km	12,89 km	+35,7 m.
BDCARTO	12,62 km	12,73 km	+112.4 m.

Table 2. Estimation of the impact of polygonal approximation

Impact of the terrain

The BDTOPO® road network is provided with altimetry, what is not the case for the BDCARTO®. In order to assign altitudes for each objects vertex of BDCARTO®, the BDALTI® is used. Computation of lengths using altitudes provides important differences, in comparison with a simple 2D computation, as shown in Table 3.

Dataset	Length 2D	Length 2D5	Difference
BDTOPO	12,86 km	12,95 km	+89,8 m.
BDCARTO	12,62 km	12,70 km	+80,6 m.

Table 3. Estimation of the impact of terrain

6.3 Discussion

The last contribution to consider in the length computation error is the one provided by the generalisation process. Detecting generalisation in a dataset is an ongoing task. This contribution to the final error is more complex to model, as it provides different effects on the objects shapes (such as simplification, enlargement of curves, bends removal).

Nevertheless, we can compute the corrected distance of the road D112 on both BDTOPO® and BDCARTO® datasets (Table 4). We suppose we can aggregate the errors modelled previously.

Dataset	Total Error	Length Min	Length Max
BDTOPO	134.1m.(+/-14,7)	12,97 km	13.00 km
BDCARTO	201.4m.(+/-92,4)	12.73 km	12.91 km

Table 4. Computation of the corrected maximum and minimum lengths by addition of errors

For the BDTOPO®, which it is not a generalised dataset, the addition gives a corrected distance of 12,99 km (+/-14,7 m.) where the most important part of the error is provided by not taking account of the terrain. For the BDCARTO®, the addition of errors gives a corrected distance of 12,82 km (+/- 92,4 m.). If we use the maximum value of the corrected length (12,91 m., which is close to the corrected length of the BDTOPO®), the error reaches 300 m, with an important impact of the polygonal approximation of curves.

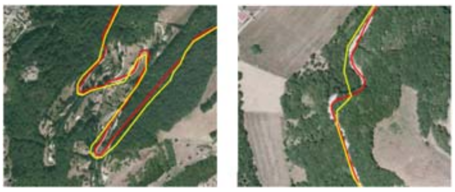

Figure 10. Example of generalisation impact on BDCARTO® road network (in yellow)

As we know that the BDCARTO® is captured using generalised 1:50000 IGN maps, as exposed in Figure 10, the impact of generalisation can be important and have to be estimated.

Thus, the example of the D112 well illustrate that the different components of error provide different impacts on the computation of length. This road has been voluntarily chosen because of its mountainous configuration, which exaggerates impacts of the terrain or also generalisation. In comparison, the same computation of errors performed in a region of plain provides results significantly different. For example, for a road of 2,96 km, the total error is 2,5 m (addition of impacts of projection system, polygonal approximation and terrain), with a digitizing error uncertainty of +/-7,18 m. This result illustrates that the impact of contributions of the final length error is completely different according to the geographical context.

Nevertheless, estimations performed show that integrating knowledge on production processes can help to understand the components of the error in a dataset and to estimate their impact in the length computation error.

7. CONCLUSION AND PERSPECTIVES

This paper presents an overview of an ongoing research on the conception of a model to evaluate geometric imprecision impact on classical measurements, and its communication to the final user. The integration of knowledge on production processes constitutes the original aspect of this work as we assume it provides variable contributions to the final error according to the geographical context, impacting the computation of length and area. Experimentation performed on a road network illustrates the respective impact of each production process in the length measurement error. In perspectives, the elaboration of the model will suppose to integrate measurements, contextual and shape indicators with additional information in order to constitute a knowledge base. Validation of hypothesis and rule base represents the core of the model, as far as the understanding of the combination of production processes impact in the final error. Finally, the communication of results constitutes the ultimate step to attend.

REFERENCES

Bel Hadj Ali, A., 2001. Qualité géométrique des entités géographiques surfaciques, Application à l'appariement et définition d'une typologie des écarts géométriques, PhD Thesis., Marne-la-Vallée University, France, 210 pp.

Bonin, O., 2002, Modèle d'erreur dans une base de données géographiques et grandes déviations pour des dommes pondérées ; application à l'estimation d'erreurs sur un temps de parcours, PhD Thesis, Paris 6 University, France, 147 pp.

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), 8-10 July, Nantes (France).

Burrough, P., 1986. Principles of Geographical information system for Land Ressources assessment. Oxford University Press, 193 pp.

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

Devillers, R., 2004. Conception d'un système multidimensionnel d'information sur la qualité des données géographiques, PhD Thesis, Laval University, Québec, Canada and Marne-la-Vallée University, France, 157 pp.

Devillers, R., and Jeansoulin, R., 2006. *Fundamentals of Spatial Data Quality.* ISTE, London, 312 pp.

Fisher, P., 1999, Models of uncertainty in spatial data. In: *Geographical information systems: principles, techniques, management and applications*, John Wiley and sons, London, Vol. 1, pp. 191-205

Goodchild, M., Sun, G. and Yang S., 1992. Development and test of an error model for categorical data. *International journal of geographical information system*, 6(2), pp. 87-103

Goodchild, M., 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research, 2*, pp. 24-32

Heuvelink, G., 1998, Error Propagation model in environmental modelling with GIS. Taylor and Francis, London, 127 pp.

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), pp. 11-21

Kresse, W. and Fadaie, K., 2004. *ISO standards for geographic information.* Springer, Berlin, 322 pp.

Mustière S., Devogele T., 2008. Matching networks with different levels of detail, *GeoInformatica*, 12 (4), pp. 435-453

Olteanu A., 2008. A Multi criteria fusion approach for geographical data. In *Quality Aspects in Spatial Data Mining*, Taylor and Francis, pp 45-56

Oort, P., 2005. *Spatial Data Quality: from description to application*, PhD Thesis, Wageningen University, The Nederlands, 132 pp.

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

Vauglin F., 1997, Modèles statistiques des imprécisions géométriques des objets géographiques linéaires, PhD Thesis, Marne-la-Vallée University, France, 286 pp.

Worboys, M., 1998. Computation with imprecise geospatial data. *Computers, Environment and Urban Systems* 22(2), pp. 85-106

Worboys, M.F., 1998. Imprecision in finite resolution spatial data. *Geoinformatica,* 2(3), pp. 257-280

Ying He, 2008. *Spatial Data Quality Management*, PhD Thesis, University of New South Wales, Sydney, Australia, 188 pp