Automated Linear Geometric Conflation for Spatial Data Warehouse Integration Process

Lionel Savary, Karine Zeitouni

PRiSM laboratory, 45 Avenue des Etats-Unis, 78035 Versailles
France
{Lionel.Savary, Karine.Zeitouni}@prism.uvsq.fr

SUMMARY

In spatial data warehouses, the quality of data greatly depends on the integration and cleaning process during which the data warehouse is fed. In this paper, we present a new geometric conflation algorithm for linear data type which gives better quality results than existing algorithm and thus, improves the quality of spatial data warehouses. While most of existing methods focus either on metric distance or shape to match spatial data, our algorithm takes into account these two factors. Consequently it resolves many complex cases which are not resolved by existing methods, and limits the need of expert intervention. In fact, the returned result, using the proposed algorithm, is close to human visual intuition. Experimental results show the effectiveness of our algorithm.

KEYWORDS: *Conflation, Data Integration, Quality, Spatial data warehouse*

INTRODUCTION

Geographic data type describing real world phenomena are generally stored in heterogeneous and distributed sources: files, databases, Geographic Information System (GIS). A geographical entity is defined by its spatial (geometry) and its non-spatial (semantic description) components. The sources where data are stored are, in most cases, disparate according to sources specificities, making complex their access and their analysis. These disparities appear in the data format, the geometric representation, the accuracy depending on target application focus, and the query language of sources. In conventional information systems, these problems are usually dealt with data warehouse tools where the data are collected, integrated and historized from different operational systems.

In this paper, we focus on the integration of geographical data. When data come from heterogeneous sources, several types of inconsistencies can arise, as well in their spatial as in their non-spatial components (Branki & al, 1998) (Savary & al., 2003) (Zhang & al., 2000). Conflicts in spatial data are much more complex to resolve than conflicts in non-spatial data (Devogele & al., 1998) (Park, 2001). They generally relate to the granularity (scale), resolution, accuracy, raster or vector format, representation model, etc. Indeed, data sources represent different point of view on the same entities depending on the focus of the application. For instance, the road network is represented differently for navigation purpose (as GPS navigation tools), for global traffic assessment or for road maintenance as the management of public works. The difficulty is to link entities that have been represented with different point of views.

This is the case in the framework of the European project HEARTS[1], the data warehouse contains, amongst other things, information coming from different sources on the road network. This warehouse should integrate, on the one hand, the road network describing in detail the junctions and

---

the links' geometries, and on the other hand, the traffic layer that provides a graph oriented point of view and does not represent the links geometry. Although these two layers provide a description of Lille's (town in France) road system, they were created at different periods with distinct scales and level of details. Precisely, the road network geometry is represented by a polyline (i.e. a line string) while traffic network geometry is described by a single line segment. Moreover, the relationships between these two entities are not explicitly defined (they have no common identifier). Therefore, they require geometric conflation tools (Devogele, 2002) (Fréchet, 1906) (Gabay & al., 1994) (Hangouet, 1995) for discovering and establishing links between them. In our case, one must consider not only the distance but also the shape similarity between two entities. In this paper, we only focus in the study of linear geometric conflation.

The rest of the paper is organized as follows: the second section presents related works; then, the third section introduces the problem specificities; the fourth section underlines the limits of existing works; the fifth section describes the proposed algorithm; the sixth section gives the experimental results of our approach and shows its effectiveness on real datasets; finally, a general conclusion summarizes our contribution and the main advantages of the proposed algorithm.

GEOMETRIC CONFLATION METHODS FOR LINEAR OBJECTS

Conflation is a process that consists in establishing links between geographical objects representing real world phenomena. Some geographical data have multiple representations of their location depending on the source and/or the initial objective when they were produced. In the literature, we divide four categories of conflation methods:

- Those that make delimitation of a conflation zone.
- Those defining distance measure between objects.
- Those based on shape similarity .
- Those based on shape similarity and distance measure.

Conflation zone

There exist different techniques based on the delimitation of conflation zone. One technique is the buffer method where a distance d is defined and associated to a geometric object x. Each object whose distance compared to the object x is under d can be matched with x. The second technique is *epsilon band* method where a tolerance zone is associated to points and segments composing the polylines. In this method, a circle of tolerance is associated with each point whose ray changed according to the nature of the represented point. Then, the circles associated with each extremity of the segment are linked by their common tangent in order to build the *tolerance band* (Gabay & al., 1994).

Distance between linear objects

Through this technique, two objects of corresponding classes are joined if the selected distance is under a specified threshold. In the literature, many distances have been defined for discovering relations between linear objects. Among these distances, one can distinguish the average distance (McMaster, 1986), where the computation of distance between two objects is carried out by dividing the total surface of displacement by the size of the reference segment. The Hausdorff distance computes the maximal gap between two segments composing the polylines. By definition, two segments $L1$ and $L2$ are at a Hausdorff distance less than d units from each other if each point of $L1$ is at less than d units from at least one point of $L1$ (Hausdorff, 1919) (Mustière, 1995). Fréchet distance (Fréchet, 1906) only computes the couple of points that should be corresponding visually, which is not the case in Hausdorff distance. Brown (Brown & al., 1995) explored the use of existing GIS functions such as rubber sheeting and dynamic segmentation to conflate network data. Filin (Filin and Doytsher, 2000) developed a projection based strategy. First the breakpoints of the two polylines are identified and matched. Then, the segment projections (translation) are applied, using the normalized

cumulative distance. Devogele (Devogele, 2002) proposes an algorithm based on the Fréchet distance in order to make a linear conflation between polylines, by using a distance matrix pin-pointing the distances between nodes of polylines to compare. It allows matching pairs of nodes of those polylines by computing the shortest distance between nodes.

Shape similarity

According to (McMaster, 1986) (Mustière, 1995), the different criteria for linear object conflation according to their shape are: the proportion of edges size, the proportion of turning round average size, the difference between direction of each segment, the proportion of the number of intermediate points, the proportion of angles sum between segments, the proportion of turning round number. These different criteria allow comparison of linear objects only through their shape.

Shape and distance similarities

Xiong (Xiong & Sperling, 2004) developed a three-stage network matching procedure based on: node matching, edge matching and segment matching. Edge matching is defined as the cumulative measure of the average difference angle between the two edges to be matched, and their average distance. In the same spirit, Pendyala (Pendyala, 2002) proposes an algorithm based on two steps: the bottom-up and the top-down computation. The bottom-up step starts with node matching, then, proceeds on segment matching, and finally ends up with edge matching. The top-down procedure is the reverse one. The bottom-up computation will find correspondences where node matches can be established, while the top-down computation will find correspondences where node matches fail or network structures differ.

PROBLEM STATEMENT

We define N:M type conflation as the matching between two polylines, and 1:N type conflation as the matching between one straight line and a polyline composed of N lines. In many cases, as in our application, there is a need to conflate a graph oriented point of view (*partial and simplified network*) and a *detailed network*. Let E_T denotes a *simplified network* where each link is grossly described by a segment line, and E_P the *detailed network* composed of polylines. Hence, the integration problem here is related to the integration of *detailed network* objects $R \in E_P$ and objects belonging to a partial and *simplified network* ($T \in E_T$). The matching process between two entities T and R raises the following problems:

- How to determine potential R candidate associated to a given T?
- How to choose the right candidate which corresponds to visual judgment given by an expert?

LIMITS OF EXISTING WORKS

Existing approaches allow N:M type conflation but are not well suited for 1:N type conflation. In some simple cases as shown in figure 1, existing algorithms may give the correct result. Here R1 (R1 $\in E_P$) will be chosen as the best matching candidate with T, which can be verified visually. However, in complex cases as in figure 1, the result may be wrong, or some methods require the intervention of experts to visually choose the right matching. Although Xiong (Xiong & Sperling, 2004) and Pendyala (Pendyala, 2002) take into account both shape and distance, they do not resolve the case in figure 2. Indeed, the cumulative distance between T and RR2 could be less or equal to this distance between T and RR1. Thus, RR2 (RR2 $\in E_P$) will be chosen as the best matching candidate with T $\in E_T$. But, visually, RR1 (RR1 $\in E_P$) appears to be a better candidate than RR2 although it is further away from T.
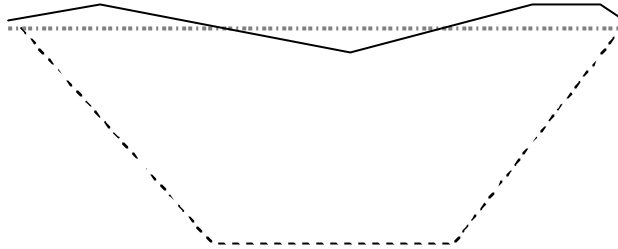
Figure 1: case 1 of conflation
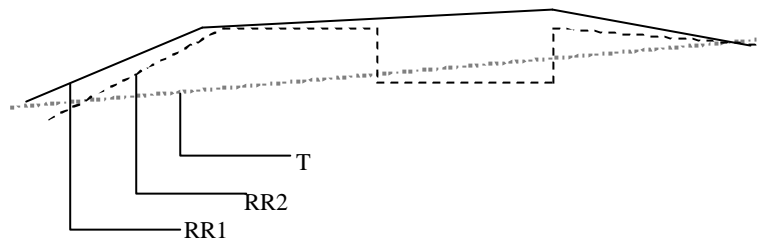
Figure 2: case 2 of conflation

Our proposed algorithm resolves these different cases of conflation and gives better quality result, close to human eyes and experts' opinion. The proposed method not only takes into account the distance between objects but also their respective shapes, and allows affecting a weight more or less important between these two criteria thanks to a specific distance defined below.

AUTOMATIC CONFLATION ALGORITHM

The proposed general algorithm is based on a succession of filtering and smoothing process, which allows reducing the number of potential candidates before the distance computation. These preliminary phases are essential both for improving the quality of the result and for considerably optimizing the performance in term of processing time. Indeed, without filtering, the distance computation would be very expensive on the totality of the the detailed roa network. Here is the chaining up of the five steps of the proposed algorithm:

1. Semantic filtering.
2. Nodes matching and searching delimitation space.
3. Polylines smoothing.
4. Candidate selection using an appropriate SM (Shape-Metric) measure.

Steps 1 to 3 are pre-processing steps. Step 1 allows reducing the number of candidates. Step 2, delimitates the search space. Step 3 reduces the number of angles in a polylines by a smoothing process. Finally, step 4 provides a new measure for better matching candidates, by combining shape and metric distance.

Semantic filtering

This phase may be useful to filter irrelevant entities using pre existing knowledge on the data semantic. For instance, the experts of transport models exclude secondary axes during the construction of the traffic network map. In our case, we have used an attribute of the (detailed) road network related to the traffic class in order to select network objects whose traffic class is upper than 1500 vehicles per day. This leads to discard the following entities: secondary roads, non circulated paths, private paths, rural path, etc.

Nodes matching and searching delimitation space

In this step, local perimeters (Cd and Cf) are defined around each node of T. Then, all polylines {R1, R2, R3} $\in E_P^3$ starting inside one perimeter and ending in the other are taken into account. A buffer Bf of size h = L/k, (where L denotes the length of T and k a real) is constructed in order to reduce the number of candidates. Thus only R1 and R2 are taken into account (see figure 3).
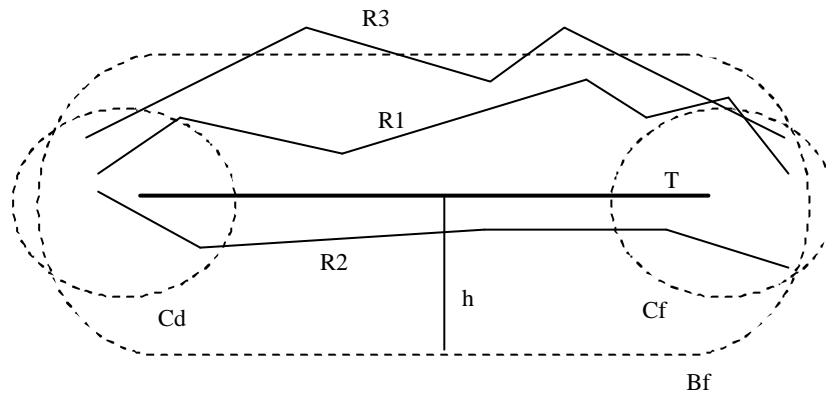

Figure 3: Nodes and path filtering.

Polylines smoothing

Inspired from geometry generalization, this process simplifies the geometry when the angle is close to 180°. As in figure 4, each nodes of R2 whom angles (a, ß, f) are greater than a specified threshold are removed (ß). Then all other nodes angles are re-computed (a',f '), and the polyline R2' is obtained.
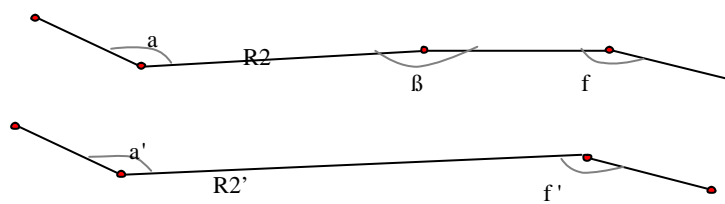

Figure 4: Polylines smoothing.

The SM (Shape -Metric) distance

Once the filtering and smoothing process are done, we propose a new distance measure. This new measure is a combination of a metric and shape distance and is formally defined as follows.

Metric distance

The Euclidean distance is used to compute distance (dn_i) between each nod (n_i) of $R \in E_P$ and $T \in E_T$ ($1 = i = 5$, in figure 5). Also, the distance (d_d and d_f) between each extremity of the two objects are computed.
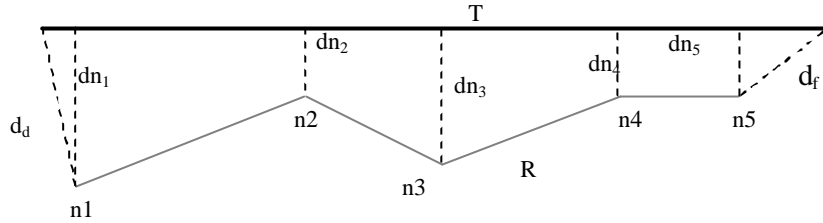

Figure 5: Metric distance.

Then the corresponding average distance $d_{T,R}$ is:

$d_{T,R} = ((\Sigma_i dn_i) + (d_d + d_f))/ (K+2)$. With $i \in [1..K]$, K = number of nodes in the polylines R, dn_i the distance between the ith node of R with T, d_d and d_f the distances between respectively the first and the last node of each entity (see figure 5).

Shape distance

The shape distance between each entity is represented by the average angle measure between each line of R and T (see figure 6).
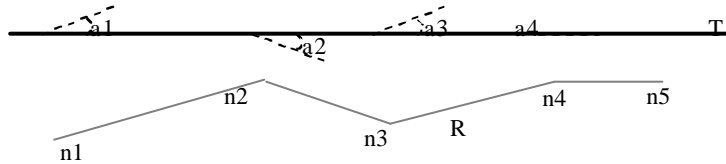

Figure 6: Shape distance.

Figure 6 shows the angles (a1, a2, a3, a4) between each line of R starting from nodes (n1, n2, n3, n4) with T. This average distance $ß_{T,R}$ is defined by:

$ß_{T,R} = (\Sigma i\ ai) / K$ where ai stand for the ith node of R and K, the total number of angles composing all candidates.

The SM distance

In order to compare these two distances (metric and shape), we should normalize the measure unit as a number between 0% and 100%, for shape similarity and metric distance Let values D_h and D_a be respectively the average similarity of metric distance and shape between these two entities:

- $D_h = (d_{T,R} * 100) / h$ where h corresponds to the gap between T and the buffer.
- $D_a = (ß_{T,R} * 100) / p$ where p stands for to the flat angle corresponding to the extreme value of dissimilarity between R and T.

Thus, the similarity between two objects could be expressed by a normalized couple measures called Shape-Metric (SM for abbreviation) distance: $D_{SM} = (D_h , D_a)$. By drawing all SM distances onto a

2D space will reflect in fact the *global similarity* percentage of the target objects. In the example of figure 7, we have defined a two dimensional referential: D_h (ordinate) et D_a (abscise).

Figure 7 represents three SM distances $D_{SM3} = (20, 30)$, $D_{SM4} = (30, 20)$, $D_{SM5} = (40, 9)$ of {R3, R4, R5} $\in E_P{}^3$ with a given traffic object T $\in E_T$. Here we can notice that:

- R3 has a closest metric distance than R4 and R5 to T, but the furthest shape distance.
- R4 has a more similar shape to T than R3 and a closest metric distance to T than R5.
- R5 has a closest shape to T than R4 and R3, but a furthest metric distance.

The user or expert must choose the proportion weight between the shape and metric distance, in other words, the importance given to the metric distance proportionally to the shape. Let P be the value standing for this proportion: $P = Vd/Vf$ where Vd and Vf are integer values standing respectively for the distance and the shape weights. For example $P = 2$ means that twice more weight is given to the distance than to the shape.

Now, in order to compare these distances and affect more weight to the shape or metric distance, the user has just to define a sliding straight line of slope -P corresponding to an equation: $D_h = -P*D_a + b$ (where P stands for the weight and b moving with the sliding motion of the straight line from 0 to max (D_h)). Then, all points located on the left side of this sliding straight line will be considered as closest to T than those located on her right side.

For example, with $P = 2$ in figure 7, sliding d1 from the origin point (0, 0) to the right, we encounter first D_{SM3}, which means R3 matches better with T. If $P = 1/2$, sliding d2 will intersect first D_{SM5}, and then, R5 matches better with T. This graphic representation shows how one can adjust and adapt the choose of P proportion between shape and metric weights.
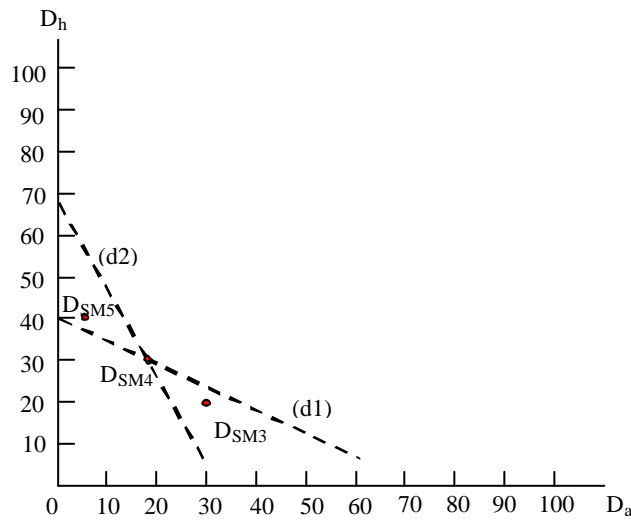
Figure 7: Distances referential.
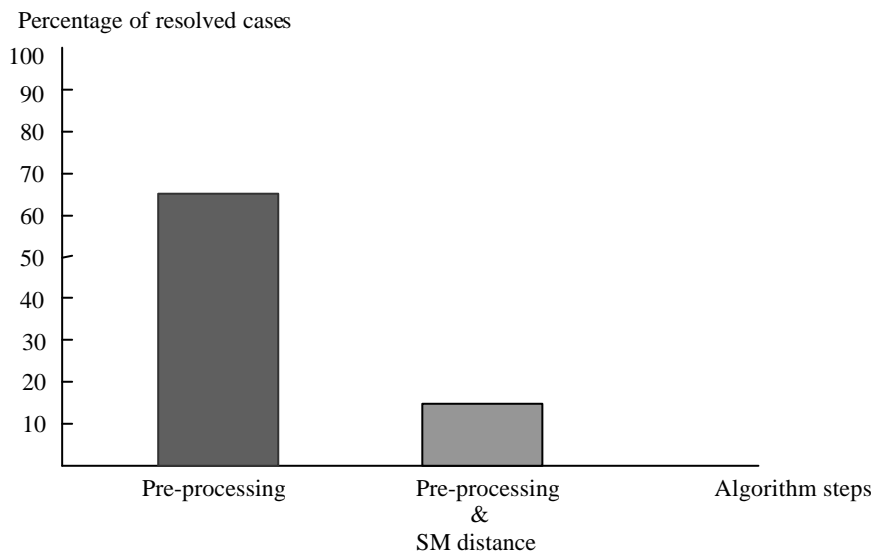
EXPERIMENTAL RESULTS

This algorithm has been implemented on real data: *traffic* (simplified network) and *road network* (detailed network) of Lille French town. These data are composed of 36709 *road network* polylines and 4308 *traffic* straight lines.
For the experimentations:

1. The weight *P* has been fixed to 1/3. This value has been chosen empirically, in order to resolve most conflation cases, including complex cases like the examples in figures 2.
2. The parameter *k* used to compute the size of the buffer Bf (figure 3) has been fixed to 2. Thus the size of the buffer is half of the length of a *traffic* object.
3. The radius of the local perimeters Cd and Cf has been empirically fixed to fifth of the *traffic* object length.

The results are described in figure 8:

- After the two first steps of the pre-processing, the algorithm finds only one road object for 2874 *traffic* objects. This represents 67% of *traffic* objects that match with *road* objects.
- For the remaining *traffic* objects that match with several *network* objects, 626 (15%) traffic objects are resolved with the proposed SM distance. These correspond to complex cases (figures 1 and 2).
- The rest (18% of cases) are either noise data corresponding to bad quality that can not be processed as well by the algorithm as by a human expert, or some rare specific cases such as the general rules do not apply. This is the only cases that should be checked manually.

Percentage of resolved cases

Figure 8: Experimentation results.

DISCUSSION

The algorithm of Devogele (Devogele, 2002) is based on a distance matrix. The shortest distance between each node of objects determines the best matching candidate. In our case, since the *traffic* objects are represented as a straight line, only the starting and the ending nodes will be used as a criterion to match with *road network* objects. Thus in the example of figure 9, this algorithm will find R2 as the best matching *road network* objects for the *traffic* object T. This is because the starting and ending nodes of R2 are nearest than the starting and ending nodes of R1, although R1 is in reality a better choice.
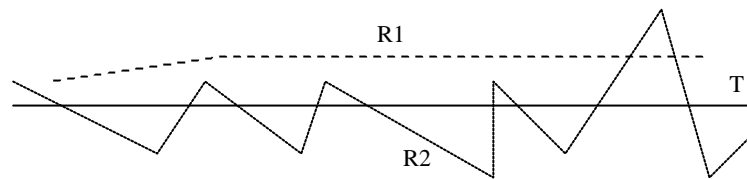
Figure 9: Example of conflation.

The methods of Xiong (Xiong & Sperling, 2004) and Pendyala (Pendyala, 2002) take into account both shape and distance, using a cumulative distance. As emphasized in the fourth section above, this do not resolve complex matching problems as in our application.

Consequently, related works in linear geometric conflation are not well suitable to 1:N conflation type. The returned results do not correspond to human perception, making manual interventions necessary to resolve these cases. Our proposed algorithm, allows reducing human intervention and give better results.

CONCLUSION

Data quality of geographical data warehouse greatly depends in the integration process. This paper has presented a so-called 1:N new linear geometric conflation algorithm. The proposed algorithm takes into account both shape similarity and distance measure for a better matching process. The returned result is close to human visual intuition. It automatically resolves complex cases, which usually need expert intervention in existing works. Thus it gives better result and improves geographical data quality and interoperability in spatial data warehouse integration process.

BIBLIOGRAPHY

Branki T., Defude B., Data and Metadata: Two-Dimensional Integration of Heterogeneous Spatial Databases. In Spatial Data Handling 98 Conference Proceeding, pages 172-179. Vancouver, BC, Canada, July 1998;

Brown J., Rao A., Baran J., Automated GIS conflation: coverage update problems and solutions. Proc. Of Geographic Information Systems for Transportation Symposium (GIS-T). American Associantion of State Highway and Transportation Officials, Sparks, Nevada, Pages 220-229.

Devogele T., Parent C., Spaccapietra S., On Spatial Database Integration. International Journal of Geographic Information Systems, Special Issue on System Integration, Vol. 12, No 3, 1998.

Devogele T., A new Merging Process for Data Integration Based on The Discrete Fréchet Distance. 10th International Symposium on Spatial Data Hangling. pages 167-181. ISBN 3-540-43802-5 Springer-Verlag. Ottawa, Canada, July 9-12, 2002.

Fréchet M., Sur quelques points du calcul fonctionnel. Rendiconti del Circolo Mathematico di Palermo, 22 :1-74. 1906.

Filin S., Doytsher Y., A linear conflation approach for the integration of photogrammetric information and GIS data. IAPRS, Vol. XXXIII, Part 33/1, pages 282-288, Amsterdam, 2000.

Gabay, Yair, Yerahmiel Doytsher. Automatic Adjustment of Line Maps. In proceedings of GIS/LIS 94. Bethesda: ACSM-ASPRS-AAG-URISA-AM/FM, 1:332-340. 1994.

Hangouet J.F, Computation of the Hausdorff distance between plane vector polylines. Twelfth International Symposium on Computer- Assisted Cartography. vol. 4, pages 1- 10. 1995.

Hausdorff F., Dimension und ausseres, Mass. Mathematische Annalen, num. 79, pages 157-179. 1919.

Jarke M., Lenzerini M., Vassiliou Y., Vassiliadis P., Fundamentals of Data Warehouses. Springer edition, 2000.

McMaster R., A statistical Analysis of Mathematical Measures for Linear Simplification. The American Cartographer, vol. 23. 1986.

Mustière S., Measure of linear generalisation quality. DESS Cartographie satge repport, Paris I University, COGIT. 1995.

Park J., Schema integration methodology and Toolkit for Heterogeneous and Distributed Geographic Databaases. In Journal of the Korea Industrial Information Systems Society, V6, pages 51-64. 3 September 2001.

Pendyala R.M., Development of GIS-Based Conflation Tools for Data Integration and Matching. Final Report: Executive Summary. Research Center, Florida Department of Transportation, 605 Suwannee Street, MS 30 Tallahassee, FL 32399-0450, 2002.

Savary L., Zeitouni K., Spatial Data Warehouse – A Prototype. In proceedings of the Second EGOV International Conference. LNCS 2739, pages 335-340. Prague, Czech Republic, September 2003.

Savary L., Wan T., Zeitouni K., Spatio-temporal data warehouse design for human activity pattern, GIM , Zaragoza, Spain, 30 August – 3 September 2004.

Zhang J., Javed M., Shahee A., Gruenwald L., Prototype for Wrapping and Visualizing GeoReferenced Data in a Distributed Environment Using XML Technology . In Eighth International Symposium of ACMOS, McLean, Virginia, November 2000.

Xiong D., Sperling J., Semiautomated matching for network database integration. ISPRS Journal of Photogrammetry & Remote Sensing 59. Pages 35-46. 2004.