A MULTI-CRITERIA FUSION APPROACH FOR GEOGRAPHICAL DATA MATCHING

A-M Olteanu

COGIT Laboratory, Institut Géographique National
2 av. Pasteur, 94160 Saint-Mandé, France
Tel.: + 33143988543; Fax: + 33143988581
ana-maria.olteanu@ign.fr

KEY WORDS: data matching, data fusion, imprecision, uncertainty, The Theory of Evidence

ABSTRACT:

Currently, there is a multiplicity of geographical information to describe the same reality. Nevertheless, we realise that there is independence between databases that represent the same reality and this affects both data users and producers. Integration seems to be the issue to that problem. One of the steps of the integration process is the data matching process that furthermore represents our issue. The purpose of this paper is to define a data matching approach based on the Theory of Evidence. The goal is to model explicitly the imperfection and to fusion many criteria of data matching in order to improve the quality of the results. Our approach is carried out in threes steps, consisting in computing the basic belief masses and fusioning criteria and assumptions. The basic belief assignments for each source are modelled in order to allow to weight a possible matching and to weaken it, in the contrary case. Fusion process based on combination of distinct sources may generate conflict. In this paper, different operator are tested in order to manage conflict and we show that the choice of an appropriate operator is crucial.

1. INTRODUCTION

At present, there is a multiplicity of geographical databases (GDB) to describe the same reality at different spatial and semantic resolutions. Many applications, like quality evaluations, propagation of updates, incoherence detection and study of several adjacent zones require the integration of geographical databases.

The integration process (Devogele *et al.*, 1998; Sheeren, *et al.*, 2004) is usually carried out in three stages. The first one is a pre-integration stage that consists in enriching schemas and modelling standardisation to use a common model. The second stage contains two processes, schema matching and data matching which are not completely independent. Finally, the last stage called integration, is about defining an integrated schema and data specification, and actually populating the integrated GDB.

In this context, we focus on the data matching process and our aim is firstly, to study how the quality of geographical data influences it and secondly, how imperfection could be managed in order to improve it. Thus, the Theory of Evidence is used to model imperfection but also to fusion different knowledge provided by different sources in order to make a decision in the matching process.

The paper is organised as follows. In Section 2, the related work is discussed. Section 3 describes the frame of the Evidence theory. Our approach and some results using the theory of evidence in the matching process are given in Section 4. Finally, section 5 concludes the paper and gives some future issues.

2. DATA MATCHING PROCESS

2.1 Related work

Data matching process (Walter and Fritsch, 1999) is a tool that can be used to find corresponding elements in different databases.

The data matching process is used in many fields handling geographical information such as integration of geographical data (Devogele, 1997; Samal *et al.*, 2004; Volz, 2006; Mustière, 2006), automated updates propagation from one database to another (Badard, 1998; Gombosi *et al.*, 2003), quality analysis (Bel Hadj Ali, 2001) or difference and inconsistency detection between databases (Bruns and Egenhofer, 1996; Sheeren *et al.*, 2004).

Many matching methods developed in the literature have revealed good results and efficiency on certain data types in selected test areas. Matching algorithms depend on the geometry of the geographical feature to match, the topological relations, the databases scales, and last but not least, the semantic properties.

In the following section, some existing methods for matching different geographical databases are presented. The methods are differentiated according to the types of features. We distinguish two important methods: methods for isolated data, i.e. data that are relatively independent from each other and methods for networks.

In (Bel Hadj Ali, 2001) a matching algorithm proposed for polygonal data uses geometrical tools as intersection, surfacic distance, etc. (Beeri *et al*, 2004) proposed four methods based on geometrical criteria and the approach relies on a probabilistic consideration. A semantic matching is also possible for isolated data when a toponym is present in the databases. Many works, which compare string distance metrics, were proposed in the literature (Levensthein, 1965; Cohen *et al.,* 2003).

Concerning linear networks, many matching algorithms have been developed in the literature. (Walter and Fritsch, 1999)

developed a method that match features in two different databases defined at similar scales and based on geometrical and topological criteria. It relies on a statistical approach.
More generally, many matching algorithms for two network databases at similar scale (Voltz 2006; Haunert, 2005) or different scale (Devogele, 1997; Zhang *et al*., 2005; Mustière 2006) were proposed. Generally, the matching process is based on different criteria and it is carried out in several steps.
We take note that all the methods seen, either applied to ponctual, linear or polygon data, or employed to match two databases at different scales or at the same scale, use geometrical, semantic, or topological criteria. All the criteria are usually applied one by one and, more importantly, most approaches do not explicitly model data imprecision.
Therefore, our objective is to find a matching approach that takes into account all the criteria at the same time and also imperfection.

2.2 Quality of Geographical Data

Geographical data are represented at various levels of abstraction. They could be vague, inaccurate or incomplete. However, these divergences between reality and representation could be acceptable within the framework of certain applications. Quality has a major impact when a decision must be made.
There are some concepts like imprecision, uncertainty, vagueness, error, etc., that are used in the field of geographic information to measure the quality. There is no standard definition of these terms so that conflicts may appear between the definitions used in the field of geographic information and artificial intelligence (AI), but also between the ways of using these concepts by different authors in the geographic information field.
So, the concept of uncertainty is generally used in order to describe imperfect data. Many taxonomies are carried out and in most of the cases, uncertainty is the base node of the taxonomy (Devillers 2004; Fisher, 2003). The proposed taxonomies aim at describing what kind of uncertainty appears in the data when the database is instantiated. This taxonomy is guided by the distinction between uncertainties arising from well-defined objects and poorly defined objects. In those taxonomies concepts like error, vagueness, ambiguity, discord and non-specificity are employed.
In AI, three types of imperfection are distinguished (Bouchon-Meunier, 1989; Colot, 2000):

- Imprecision: it is related to the difficulty of expressing clearly and precisely the information (e.g. The surface of a forest is approximately of 3 km^2?).
- Uncertainty: it is related to a doubt about the information's validity (e.g. Is this forest really a coniferous forest?).
- Incompleteness: it refers to the absence of information (e.g. The toponym of the forest is not filled up).

Quality of data has been the core of many researches in the last few years. Therefore, imprecision must be taken into account when a decision must be made. Generally, in order to model imprecision, mathematical theories such as Fuzzy set Theory (Zadeh, 1965), Possibility Theory (Dubois and Prade, 1985) and the Theory of Evidence (Shafer, 1976) are employed.

We think that the concepts of the imperfect data in AI are sufficient for identifying all types of data imperfection that interest us in the matching process.
We also consider that the Theory of Evidence is useful in our case because on one hand, it allows to model imprecision, uncertainty and incompleteness and on the other hand, it allows to fusion knowledge and to make combined hypothesis.

3. THE FRAME OF THE THEORY OF EVIDENCE

The Theory of Evidence, also called the Dempster-Shafer theory was introduced by Dempster in 1967. His work concerns the lower and upper probability distributions. Based on Dempster's work, Shafer, (Shafer, 1976) introduced a model called Dempster-Shafer model, which is based on belief functions. The description of the Dempster-Shafer model is the object of this section.

3.1 The Frame of Discernment

Let Θ be a set of N hypotheses H_i, i= 1..N, that corresponds to the potential solution of a given problem. This set of hypotheses is called the frame of discernment being defined as follows: {

$$\Theta = \{H_1, H_2, ..., H_N\} \quad (1)$$

where N is the number of hypotheses

From the frame of discernment, let 2^Θ denote the set of all subset of Θ defined by:

$$2^\Theta = \{H_1, ..., H_N, \{H_1, H_2\}, ... \{H_1, H_2, .. H_{N-1}\}, \Theta\} \quad (2)$$

where $\{H_i, H_j\}$ = represents hypothesis that the solution of a problem is one of them, i.e. either H_i or H_j. We will call it a composite hypothesis or proposition.

The Theory of Evidence is based on the basic belief assignment (bba), i.e. a function that assigns to each assumption, $A \in 2^\Theta$, a value that represents how much a source believes in it. The bba is defined as follows:

$$\begin{aligned} & m: 2^\Theta \rightarrow [0,1], \\ & \sum_{A \subseteq \Theta} m(A) = 1, \\ & m(\varnothing) = 0 \end{aligned} \quad (3)$$

where m(A), $A \in 2^\Theta$, is called the basic belief mass (bbm)
$m(\varnothing)$ = the conflicting mass; it represents the mass of belief allocated to the empty set because sources are in conflict.

For example, if we consider that a matching process is based on geometry of the features, then the closer two features are, the

more the criterion believes that the features are homologous, and so the value of the bbm is important.
The bbm could be thought of as a probability measure, but the difference is that the mass of belief is assigned not only to the singletons hypotheses but also to any proposition of 2^{Θ}, i.e. to the composite hypotheses.
Each proposition $A \subseteq \Theta$, such as $m(A)>0$ is called a focal element of m.
From this moment, we consider only the focal elements in order to fusion information and to make a decision.

3.2 Dempster's rule of combination

The Theory of Evidence presents the advantage, among others, that it offers tools to combine several sources of information using the Dempster's operator.
Let us consider two sources S_1 and S_2. Each source supports a proposition A with a certain bbm: respectively $m_1(A)$ and $m_2(A)$. Let us denote m_{12} the bbm resulting from the combination of two sources using the Dempster's rule and that supports the same proposition A.

$$m_{12}(A) = m_1(A) \otimes m_2(A) = \frac{1}{1-m(\varnothing)} \sum_{\substack{B \cap C = A \\ B,C \subseteq \Theta}} m_1(B) m_2(C) \qquad (5)$$

where
$$m(\varnothing) = \sum_{\substack{B \cap C = \varnothing \\ B,C \subseteq \Theta}} m_1(B) m_2(C) \qquad (6)$$

Dempster's rule is commutative and associative and it supposes that sources to be combined are independent.
When sources are combined, a conflict may appear and it should be modelled. In this case, the conflict is assigned to the empty set, $m(\varnothing)$, see equation 6, and it is used by Dempster to normalise the resulting mass , m_{12}. Thus, the conflicting mass is redistributed correspondingly to the focal elements.
(Zadeh, 1986) has proved that this operator is very sensible of little variations of the mass of belief, which could influence the decision. Many authors (Smets, 1990; Yager, 1987; Dubois et Prade, 1988; Colot 2000; Royère 2002) proposed solutions to explain conflict's origins and to manage it.
Therefore, two hypotheses were made: the first one considers that the frame of discernment is exhaustive, which is called the closed world assumption, thus the conflict comes from the non-reliability of the sources. The second one, open world assumption formulated by (Smets, 1988) supposes that the sources are reliable and the conflict comes from the fact that none of the elements of Θ could be the solution. In this case, a positive basic belief mass could be given to empty set, $m(\varnothing)>0$, in opposition to the condition $m(\varnothing)=0$ established by Shafer (Shafer, 1976).

4. THE THEORY OF EVIDENCE IN A DATA MATCHING CONTEXT

Generally, the matching process consists in searching for each feature belonging to the reference database for potential candidates and then analysing them in order to determine the final result. Geographical data have imperfections (e.g. location could be imprecise, toponyms have different versions due to omission, abbreviation, substitution, etc.). Using several criteria one after the other in the matching process, errors could be propagated and the matching results erroneous. Therefore, imperfection should be taken into account in the matching process and criteria should be applied at the same time to obtain more relevant information.
The Theory of Evidence offers tools for modelling imperfection through belief functions and to fusion different knowledge through the Demster's rule of combination.
In this section, we describe a data matching approach based on the Theory of Evidence. Our approach consists in three steps, as we can see in figure 1. The first and the second step are considered as a local approach because candidates are separately analysed and the third step is considered like a global approach because candidates are analyzed together.
Each step is described in the next sections.

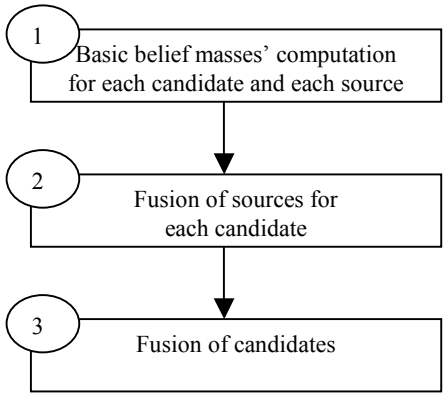

Figure 1. The three steps of our data matching approach

4.1 The frame of discernment's definition: local approach

Two databases defined by IGN France are used, that were made for different purposes, they come from different origins and does not have the same scale, i.e., a database is more detailed than the other one. The more detailed database is called comparison database and a feature that belongs to it is called comparison feature. Similarly, the less detailed database is called reference database and a feature belonging to it is called reference feature.
So, for each reference feature, we look for close features in the comparison base, which are candidates for matching. The distance that determines how far we look for candidates is an empirical one.
In our case, the number of assumptions that composed the global frame of discernment is equal to the number of candidates, i.e. each candidate could be the solution. Due to the fact that a feature may have no homologue, we introduced a new hypothesis *, standing for the assumption: the feature reference is not matched at all. So, we consider the closed world assumption, i.e. an exhaustive frame of discernment and non-reliable sources. A similar approach was presented in (Royère, 2002).
So, for each feature belonging to the reference database, RF_{ref}, we define a frame of discernment as follows:

$$RF_{ref} : \Theta = \{C_1, C_2, ..., C_N, *\} \qquad (7)$$

where N = the number of candidates,
C_i = the assumption that the homologue of the reference feature is C_i,
* = the assumption that the feature is not matched

In order to compute the basic belief assignments, we use a local approach i.e. each candidate is analysed separately, (Royère, 2002; Najjar, 2003). The aim is to find out if the candidate is the right homologue of the reference feature. Moreover, a particular case of the Theory of Evidence, called the specialised sources is used (Appriou, 1991 ; Royère, 2002). Each source specialises on a candidate and assigns a mass of belief to it. In our case, a source coincides to a criterion of data matching (toponym and geometry).
In this case we consider N_i, a subset of 2^Θ, defined as follows:

$$N_i = \{C_i, \neg C_i, \Theta\} \quad (8)$$

where C_i = hypothesis that C_i is the homologue of the reference feature,
$\neg C_i = \{C_1, C_2, ...C_{i-1}, ...C_N, *\}$ is assumption that C_i is not the homologue of the reference feature.
$\Theta = \{C_1, C_2, ...C_i, ...C_N, *\}$ is assumption that the criterion does not know if C_i is the candidate or not.

4.2 The basic belief masses' computation

In this paper, we propose two criteria in order to mach data and those criteria represent the sources in the frame of the Theory of Evidence. The first one is a geometrical criterion based on Euclidean distance and the second one is based on a string distance between the toponym of the reference feature and the toponym of the comparison feature respectively.
The two criteria are described below.

4.2.1 The geometrical criterion

The geometrical criterion is based on the Euclidean distance, d_E, between respectively the location of the reference feature location and the location of a candidate. We suppose that, the more the candidate is close to the reference feature the more the belief that the candidate is the homologue of the reference feature is high. The bba are modelled in figure 2.
In figure 2a), the selection threshold T_2 represents the distance determining how far we look for candidates. The second threshold T_1 is introduced in order to give less weight for candidates that are fairly far.

4.2.2 The toponym criterion

The second criterion that we use is based on the comparison of toponyms. A string distance d_T, between two toponyms, $toponym_1$ and $toponym_2$ is computed using the Levenshtein distance (Levenshtein, 1965) as follows:

$$d_T = 1 - \frac{d_L(toponym_1, toponym_2)}{\max(L_1, L_2)} \quad (9)$$

where d_L= Levenshtein distance, L_1= $toponym_1$'s length and L_2= $toponym_2$'s length

Note that string comparison does not take into account accents and case.

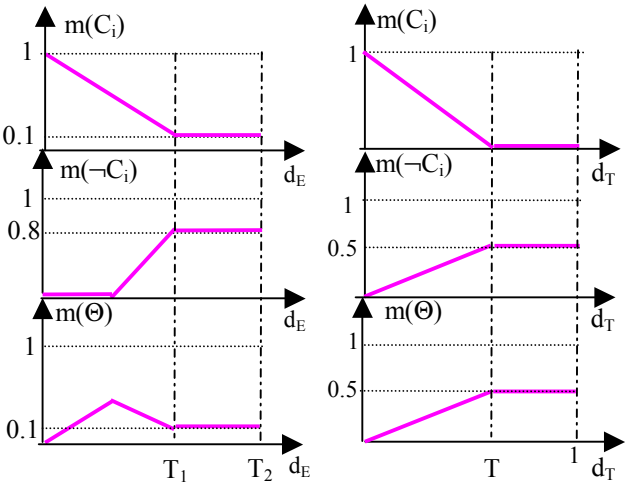

Figure 2. Modelling of the geometrical a) and toponym b) criteria.

Given two toponyms « Boulevard de Général de Gaulle » and « Boulevard de Général de Gaulle » the distance d_T is equal to 0, while the distance between « Bld du Gal de Gaulle » and « Boulevard de Général Charles de Gaulle » is equal to 0.7.
The string distance is able to accommodate minor spelling errors but it does not take into account word transposition or word substitution. Therefore, a less important confidence is assigned to it, when toponyms are different and another criterion should complement the toponym criterion.
In the figure 2 b), the bba for the toponym criterion is shown. Here, the curves are different from the geometrical criterion, in order to express that we are less confident on this source. Thus, we manage the case of ambiguity, when two toponyms which indicate the same feature are compared, but one have the official place name whereas the other have a non-official place name. It consists in decreasing the mass of belief associated with the assumption "it is not Ci the homologue of the reference feature" and increasing the mass associated with ignorance. Thus, if the distance d_T is higher than the threshold (e.g. thirty percent of letters do not resemble each other) the masses of belief assign to hypotheses Ci is not the right candidate and criterion do not know are equals to 0.5.
The next step is to combine criteria using the Dempster's rule.

4.3 The rule of combination

To combine knowledge, the following gait is followed. Firstly, each candidate is separately analysed, criteria being combined per candidate. Then, the results of the first step are combined i.e. candidates are combined in order to have a global view (Royère, 2002). In the first step we look at candidates one by one, not taken into account the others, while in the second step all candidates are taken into account in the same time.

4.3.1 The second step of the matching approach: local approach. Candidates are analysed separately, i.e. the others candidates are not taken into account (step two of figure 1). Let us consider a candidate C_i and two criteria S_1 and S_2. Each criterion assigns a bbm for each assumption as follows:

$$\begin{array}{c} m_{1,i}(C_i), m_{1,i}(\neg C_i), m_{1,i}(\Theta) \\ m_{2,i}(C_i),\ m_{2,i}(\neg C_i), m_{2,i}(\Theta) \end{array} \qquad (10)$$

where $m_{1,i}(\ )$ is the mass of belief assigned by the criterion S_1 to one of the three assumptions (C_i, $\neg C_i$, Θ) for the candidate C_i,
$m_{2,i}(\ )$ is the mass of belief assigned by criterion S_2 to one of the three assumptions (C_i, $\neg C_i$, Θ) for the candidate C_i

Sources are fusioned using the Dempster's operator of combination. Combined mass of belief for the candidate C_i, $m_{12,i}(\ )$, and conflict between the two criteria, $m_{12,i}(\varnothing)$ are obtained.

4.3.2 The third step of the matching approach: global approach.

In the second step the results of the first one are combined, (step three of figure 1). So, the results for two candidates are combined, and then these results are combined with the results for the third candidate and so on.
In order to illustrate this approach, let us consider three candidates: C_1, C_2, C_3.
In equation (11), m_{12} represents the result from the combination between C_1 and candidate C_2 and m_{123} is the result from the combination between m_{12} and candidate C_3.

$$\left.\begin{array}{l} m_{12,1}(C_1) \\ m_{12,1}(\neg C_1) \\ m_{12,1}(\Theta) \end{array}\right\} \oplus \left.\begin{array}{l} m_{12,2}(C_2) \\ m_{12,2}(\neg C_2) \\ m_{12,2}(\Theta) \end{array}\right\} \rightarrow \left.\begin{array}{l} m_{12}(C_1) \\ m_{12}(C_2) \\ m_{12}(\neg C_1) \\ m_{12}(\neg C_2) \\ m_{12}(C_3 \cup *) \\ m_{12}(\Theta) \\ m_{12}(\phi) \end{array}\right\}$$

$$\oplus \rightarrow m_{123} \qquad (11)$$

$$\left.\begin{array}{l} m_{12,3}(C_3) \\ m_{12,3}(\neg C_3) \\ m_{12,3}(\Theta) \end{array}\right\}$$

When masses of belief are combined, an important conflict may appear on the one hand, because criteria support different candidates and on the other hand because candidates are firstly analysed separately and then are analysed together. Therefore, a redistribution of conflict is necessary. Many operators that redistribute conflict exist in literature (Smets, 1990; Yager, 1987; Dubois et Prade, 1988; Colot 2000; Royère 2002). Some of them redistribute conflict locally, i.e. conflict is redistributed to assumptions that cause conflict, and others redistribute conflict globally, i.e. conflict is redistributed to union of assumptions, which caused conflict.

Some of them are associative and others are not. In consequence, the choice of operators depends on application.

4.4 Decision

In our application, a decision has generally to be taken in favor of a simple assumption (candidate). Within the context of the Transferable Belief Model, Smets defines and justifies the use of the pignistic probability decision rule (Smets 1990).

5. TEST CASES

In this section some matching results are illustrated. As we saw in section 2, our experiments concern two different geographical datasets containing punctual geographical data representing the relief. In order to illustrate our approach we chose some particular examples.

5.1 Analysis of Used Data

Two different databases of the French National Mapping Agency representing relief made from different sources and with different purposes were used in this study. They have different levels of detail. Databases are illustrated in figure 1, the less detailed named BD CARTO ®, on the left and the more detailed database named BD TOPO, ® on the right side.

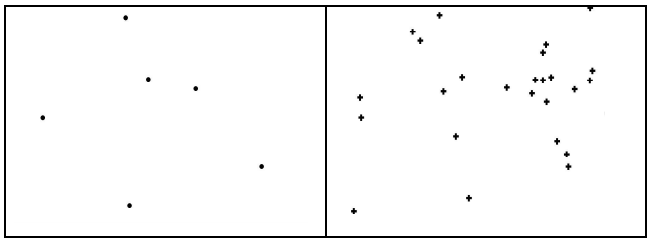

Figure 4. Relief data describing the same reality

Both databases contain information about the relief such as, mountains, mountain passes, summits, peaks, valleys, beaches, etc. These data are imprecise on the one hand by definition, i.e. the limit between valleys and mountain cannot be defined precisely, or different interpretations of concepts like summit and peak may exist.
Following our imperfection taxonomy (see Section 2.2) the next observations can be made.
First, locations of a feature are imprecise, and more, locations have various accuracy. For example a valley and a peak have not the same accuracy. Differences come from the fact that, even if both a valley and a peak are represented by a point defining usually the center respectively of the valley and peak, a valley is always more larger that a peak.
Second, the toponym is a very useful knowledge to take into account in the matching process. But toponym as well as location, also presents imperfection. For example, there are used names and official names, or the same toponym can be used for several places, or there can be various interpretations of the pronunciation, word omission, and character omission or character substitution. The toponym, may also be uncertain ("col de peyrelue **or** port vieux de sallent ") and incomplete ("col de louesque", "louesque").
Using only the geometrical criterion based on the distance between locations, errors could occur. The homologue feature

is not always the closer one. In the same way, using only the toponym criterion inconsistencies could emerge.
In order to match imprecise data, first of all the imprecision should be modelled and secondly criteria should be combined in the same time to have a global view and to avoid the propagation of the errors. There are cases when the concepts imprecision and uncertainty cannot be managed separately one from the other. For example location of a feature can be in the same time imprecise and uncertain.
Thus, we consider that an approach based on the Theory of Evidence is appropriate to our case because it allows to model imprecision, uncertainty and incompleteness. On the other hand, it offers tools to combine information, to construct simples and composites assumptions and to make a decision adapted to application, choosing either a simple assumption (e.g. a candidate) or a composite one (e.g. many candidates).

5.2 Case 1

In this case, we want to find the homologue feature for the RF_{ref} feature. After the selection step, four features are candidates. The candidate C_1 is the right homologue feature of the RF_{ref} feature. The frame of discernment is defined as follows:

$$RF_{ref} : \Theta = \{C_1, C_2, C_3, C_4, *\} \quad (12)$$

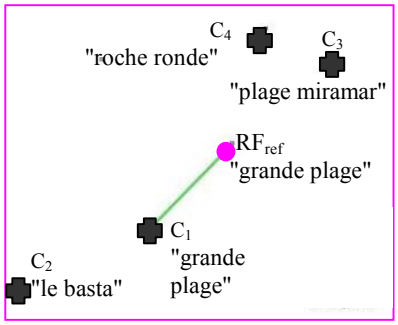

Figure 5. Data matching result

In figure 6, for each candidate, combinations of the two criteria using the Dempster's operator without normalisation are presented. After the second step, (e.g. combination of criteria for each candidate, see 4.3.1), we can see that criteria are in agreement and there is not doubt for C_1 and $\neg C_2$. Concerning the candidate C_3, criteria are in conflict, i.e. first criterion estimates that C_3 is the right homologous because it is close to the reference feature but the second criterion estimates that C_3 is not the right homologous because its toponym is different.
So, a decision could not be made at this step.

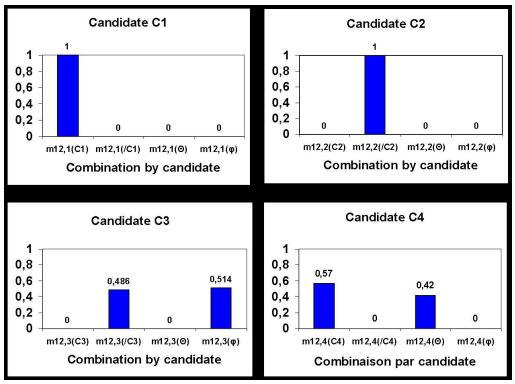

Figure 6. Combinations of the criteria per candidate

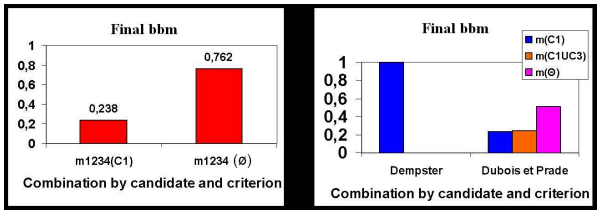

Figure 7. Criteria and candidate's combination (on the left) and conflict's redistribution (on the right)

The third step, as we presented in section 4.3.2, consist in combining all the results. In figure 7 on the left, a final bbm is presented. We see that all bbm are zero except the bbm assigned to candidate C_1 and the conflicting mass is important. Therefore, a decision cannot be made. In order to make a decision the conflict is redistributed. Many operators exist, and we decided to test two of them: the Dempster's operator and the Dubois and Prade's operator, (Dubois and Prade, 1988). The last one proposes to reallocate the conflicting mass on the union of sources that caused the conflict, whereas the first one proposes to normalise all the bbm by conflicting mass.
Applying the pignistic probability to both Dempster and Dubois and Prade's operators the solution is C_1.

5.3 Case 2

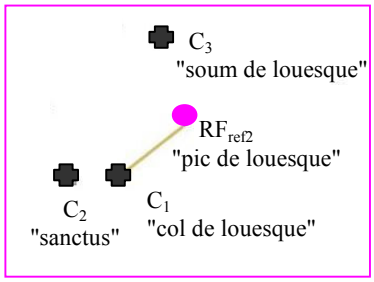

Figure 8. Data matching result

In this case, three features are candidates and decision is not easy to make because no candidate is discriminated. After the first combinations, see figure 9, we only can say that the solution is candidate C_1 or C_3, so the situation is ambiguous and no decision is obvious.

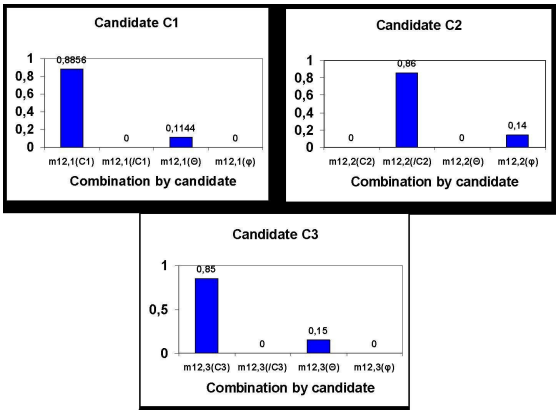

Figure 9. Combinations of the criteria per candidate

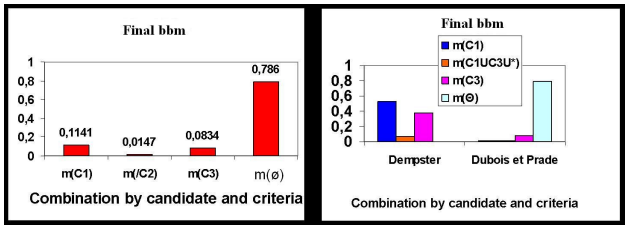

Figure 10. Criteria and candidate's combination (on the left) and conflict's redistribution (on the right)

In figure 10, on the left, candidate C_1 is discriminated but with a slight mass of belief in comparison with the mass of belief assigned to empty set, i.e. the conflict. After conflict's redistribution candidate C_1 is selected following Dempster's operator and candidate C_3, following Dubois and Prade's operator.

5.4 Case 3

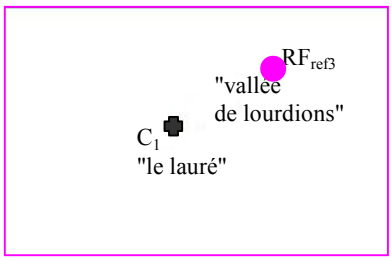

Figure 11. Data matching result

As we can see in figure 11, only one candidate is selected for the reference feature. The solution of this case is: the reference feature has no homologue.
The results are given in figure 12.

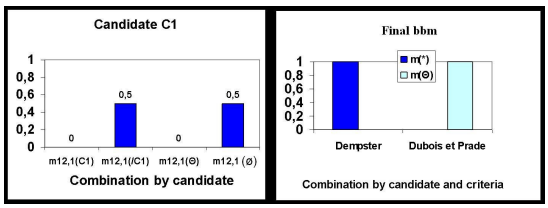

Figure 12. Criteria and candidate's combination (on the left) and conflict's redistribution (on the right)

In this case, because only one candidate is selected, the third step is not carried out. After we combined criteria, the assumption "reference feature has not homologue" seems to be the solution. Decision is ambiguous because the conflict between criteria is important. So, conflict's redistribution is necessary, see figure 12, on the right. If decision is applied after using Dempster's operator, than reference feature is not matched, otherwise, applying the Dubois and Prade's operator any decision can be made.

5.5 Conclusion of tests

The choice of an operator for managing conflict is crucial and it can lead to very different results. After watching results, we realize that, conflict may appear and that are cases when a decision cannot be make. In order to manage conflict many solutions exist such as: redefine the mass function model, added information via criteria or redistribute the conflict. The two operators that we chose are not very appropriate for our application. The Dempster's operator redistributes the conflicting mass on only one assumption, although the sources are in conflict and the conflicting mass is only used to normalise bbm, so the conflict is not really treated.
The Dubois and Prade operator treats the conflicting mass, but it is not associative. Therefore, the result depends on the order in which candidates are combined.

6. CONCLUSION

The purpose of our work is to take into account the imperfection presents in spatial data in a matching process. The results of the matching process are influenced by the data imperfection and thus it can be more or less efficient.
To undertake it, firstly, we analysed the imperfect data and the methods that can be used to manage these imperfection, both in the field of AI and spatial information. We think that an appropriate taxonomy in our case is the AI taxonomy, which classifies the imperfection in three levels: imprecision, uncertainty and incompleteness. We also consider that the Theory of the Evidence is useful in our case because it allows to model imprecision, uncertainty and incompleteness in the same time. The theory is particularly adapted because it allows to combine the sources of information and to make combined hypothesis.
Secondly, we presented a first approach to match data using the Theory of Evidence. Closed world assumption and specialized sources are supposed. Two criteria are considered, in order to compute the masses of belief.
In section 4, first results are given and we show that Dempster and Dubois and Prade's operators are not appropriate to our application, at least at this point of research, when only two criteria are used.
Finally, our forthcoming works consist to compute another associative operator that is capable to manage conflict, to improve the mass of belief function model and to carry out more tests at a large scale.

7. REFERENCES

Appriou A., 1991. Probabilités et incertitudes en fusion de données multi-senseurs. *Revue Scientifique de Technique de la Défense*, 11, pp.27-40.

Badard, T., 1998. Extraction des mises à jour dans les BDG – De l'utilisation de méthodes d'appariement. *Revue Internationale de géomatique*, 8(1-2), 1998, pp. 121–147.

Bel Hadj Ali A., 2001, Qualité géométrique des entités géographiques surfaciques –Application à l'appariement et définition d'une typologie des écarts géométriques. PhD dissertation, University of Marne la vallée.

Beeri, C., Kanza, Y., Safra, E. and Sagiv, Y., 2004. Object Fusion in Geographic Information Systems. In *Proceedings of the 30th VLDB Conference*, Toronto, Canada.

Bouchon-Meunier, B., 1989. On the management of uncertainty, In *Encyclopedia of Computer Science and Techology*, Marcel Dekker, New York, 20(5), pp 327-337.

Bruns, H.T. and Egebhofer M, 1996. Similarity of Spatial Scenes", In *Seventh International Symposium on Spatial Data Handling, Delft*. Netherlands, August, London, pp.173-184, Taylor & Francis.

Cohen, W.W., Ravikumar P. and Fienberg, S.E., 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of the IJCAI*, 9-10 August, Acapulco, Mexico, pp. 73-78.

Colot, O., 2000. Systèmes de perception d'informations incertaines-Application au diagnostic médical. HDR dissertation, University of Rouen.

Devogele, T., Parent, C. and Spaccapietra, S., 1998, "On spatial database integration", In *International Journal of Geographical Information Science*, 12(4), 1998, pp. 335-352

Devillers, R., 2004. Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales. PhD thesis, University of Marne la Vallée.

Dubois, D., and Prade, H., 1985. Théorie des possibilités: applications à la représentation des connaissances en informatique, Masson, Paris.

Dubois, D. and Prade, H., 1988. Representation and combination of uncertainty with belief functions and possibility measures. In *Computer Intelligence*, 4, pp. 244-264.

Gomboši, M., Žalik, B and Krivograd, S., 2003. Comparing two sets of polygons, In *International Journal of Geographical Information Science*, 17 (5), pp.431-443.

El Najjar M.E., 2003. Localisation dynamique d'un véhicule sur une carte routière numérique pour l'assistance à la conduite, PhD Thesis, University of Compiègne.

Fisher P. F., 2003. Models of uncertainty in spatial data. In *Geographical Information System*, 1, Second Edition, pp. 191-203.

Haunert, J.H, 2005. Link based Conflation of Geographic Datasets. In *8th ICA Workshop on Generalisation and Multiple Representation*, Coruna, 7-8 July.

Levenshtein, V., 1965. Binary codes capable of correcting deletions, insertions and reversals. In *Doklady Akademii Nauk SSSR*, 4 (163), pp.845-848.

Mustière, S., 2006. Results of Experiments on Automated Matching of Networks at Different Scales. In *ISPRS Workshop, Multiple representation and interoperability of spatial data*, Hanover, Germany, 22-24 February, pp. 92-100.

Royère C., 2002. Contribution à la résolution du conflit dans le cadre de la théorie de l'évidence :Application à la perception et à la localisation des véhicules intelligents. PhD Thesis, University of Compiègne.

Samal A., Seth S., et Cueto K., 2004. A feature-based approach to conflation of geospatial sources. In *International Journal of Geographical Information Science*, 18(5), pp. 459-489.

Shafer G., 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.

Sherren, D., Mustiere, S., Zucker, J-D., 2004. How to Integrate Heterogeneous Spatial Databases in a Consistent Way?. In *Conference on Advanced Databeses and Information Systems(ADBIS),* Budapest, September 2004, pp. 364-378.

Smets Ph., 1988. Belief Functions.Non Standard Logics for Automated Reasoning. Smets Ph., Mamdani A., Dubois D. and Prade H. (Editors), Academic Press, London, pp. 253-286.

Smets Ph., 1990. The Combination of Evidence in the Transferable Belief Model, In *IEEE Trans*. PAMI 12, pp. 447-458.

Voltz, S., 2006. An Iterative Approach for Matching Multiple Representations of Street Data. In *ISPRS Workshop, Multiple representation and interoperability of spatial data*, Hanover, Germany, 22-24 February, pp. 101-110.

Walter, V. & Fritsch, D., 1999. Matching Spatial Data Sets: Statistical Approach. In *International Journal of Geographical Information Science*, 13(5), pp. 445-473.

Zadeh, L., 1986. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. In *AI Magazine*, 7, pp. 85-90.

Zhang, M., Shi, W., and Meng, L., 2005. A generic matching algorithm for line networks of different resolutions. In *ICA Workshop on Generalisation and Multiple Representation*, Coruna, 7-8 July.

Yager R., 1987. On the Dempster –Shafer framework and new combination rules. In *Informations Sciences*, 41, pp. 93-138.