JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 10 (2015), pp. 21–46
doi:10.5311/JOSIS.2015.10.194

RESEARCH ARTICLE

Knowledge formalization for vector data matching using belief theory

Ana-Maria Olteanu-Raimond[1],
Sebastien Mustière[1], and Anne Ruas[2]

[1]Université Paris-Est, IGN, COGIT Lab, Saint-Mandé, France
[2]IFSTTAR, COSYS/LISIS Laboratory, Paris, France

Received: September 8, 2014; returned: November 11, 2014; revised: January 15, 2015; accepted: February 18, 2015.

Abstract: Nowadays geographic vector data is produced both by public and private institutions using well defined specifications or crowdsourcing via Web 2.0 mapping portals. As a result, multiple representations of the same real world objects exist, without any links between these different representations. This becomes an issue when integration, updates, or multi-level analysis needs to be performed, as well as for data quality assessment. In this paper a multi-criteria data matching approach allowing the automatic definition of links between identical features is proposed. The originality of the approach is that the process is guided by an explicit representation and fusion of knowledge from various sources. Moreover the imperfection (imprecision, uncertainty, and incompleteness) is explicitly modeled in the process. Belief theory is used to represent and fuse knowledge from different sources, to model imperfection, and make a decision. Experiments are reported on real data coming from different producers, having different scales and either representing relief (isolated points) or road networks (linear data).

Keywords: spatial data matching, vector data, fusion, imperfection, belief theory

1 Introduction

More and more spatial data representing the same "real world" co-exists. Entities are differently represented (e.g., a river may be represented by its axis or its full extent) and are intended to be used in different applications (e.g., topographic maps, route planning, risk analysis). Representations are produced using different rules by public or private organizations, or by crowdsourcing, such as with OpenStreetMap [35]. Thus datasets are independently produced and managed. This presents difficulties for data producers as well as data users when processes such as integration, data quality estimation, or updates need to be performed.

In this context a solution is to define links between features from different datasets representing the same object in the real world. This process is known in the literature as data matching [50] and is illustrated in Figure 1.

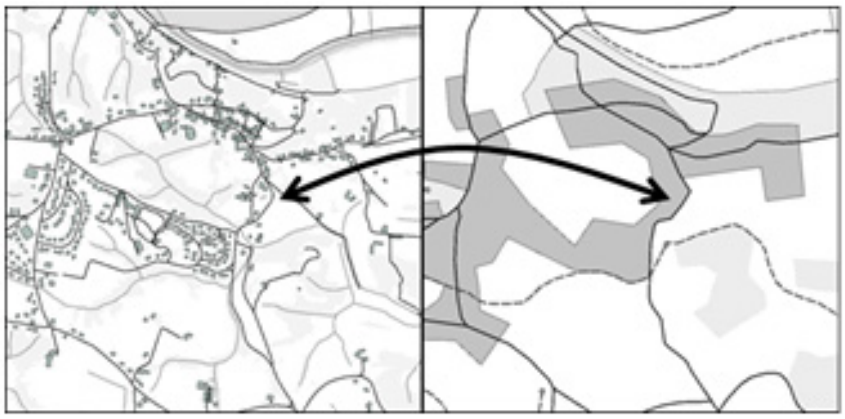


Figure 1: The matching problem. A pair of identical features in two different representations of the same world.

Data matching can be considered a tool for different purposes. Indeed, to evaluate data quality, data matching is a necessary first step to compare a given dataset to a reference one [5,22,25]. Data matching is used to identify identical features belonging to two versions of the same dataset to identify updates [23], or belonging to different datasets to propagate updates from one dataset to the other [24, 48]. Pairs of matched features are analyzed to detect inconsistencies [19,41] and for conflict detection and resolution [47]. Spatiotemporal analysis of the evolution of historical features [17,36] also requires data matching.

More generally, the integration of data coming from different sources may be required for different applications, such as for mapping or decision making. Data matching is usually a first step of both of two possible data integration approaches: multi-level—when several representations are kept separated but are linked [14, 51]; and conflation—when several representations are merged [28, 36, 39]. Data matching is not a goal per se, but a preliminary step for many applications.

This paper focuses on vector data matching, and is an extension of the approach presented in [33, 34], where it was tested on points [33] and networks [34]. In this paper, the goal is to describe in detail the knowledge which can be used for vector data matching, how it can be formalized, how different types of imperfection can be taken into account, and how this knowledge can be modeled using concepts defined in belief theory. Particular attention is paid towards showing the generality of our approach. The approach is applied to datasets having very different scales and types of representation, different types of objects, and different criteria. We compare the results of our approach with those of others.

Related works and motivations are discussed in Section 2. In Section 3, belief theory is presented. The data matching proposed approach is then described in Section 4. Before concluding, two applications of our approach are presented in Section 5.

2 Related work

In this section some related works on spatial data matching are presented and discussed from two points of view: the criteria used to compare features, and the data matching strategies to make a decision.

2.1 Data matching: A multi-criteria issue

Data matching relies on the concept of similarity, e.g., two features A and B belonging to two different datasets are matched if they are similar.

Generally, similarity measures for data matching are based on four types of information: geometric (the position and implicitly the shape); semantic (i.e., features types, such as road, summit, and building); attributes (i.e., the different characteristics of features, such as number of lanes, name, number of levels, etc.); and topological relations between features (e.g., meet, inside).

Matching processes for spatial data rely primarily on the comparison of geometric positions and relations between features. For isolated points, geometry matching can be based simply on distance. For more complex data like spatial networks, many geometric criteria can be defined, such as the comparison of orientations, lengths, levels of sinuosity. Topological relationships between features of the same dataset [31,50], or between features belonging to two different datasets [42] can also be used. These relationships are used to compare nodes (e.g., number of incidents arcs, angles between incident arcs) [7,49] or edges (e.g., orientation, length, sinuosity) [29,51], or to guide the matching sequence. For example when nodes are first matched, subsequent arcs connected to those nodes are matched [31]. For data represented by polygons, criteria based on surface distance and shape comparisons are used [3,5].

The comparison of the semantic information is usually based on either string comparison, semantic distances between concepts defined by ontologies [1, 38], or lattice approaches [16]. A detailed review on semantic measures is provided by Ballatore et al. [4]. In the literature, the approaches typically focus on only one type of geometry at a time. As a consequence, there exist specific approaches to match points [5], networks [31, 49–51] or polygons [6]. In general, for point and polygon features matching can be applied to data sets having the same or different scales. In contrast, approaches for matching networks depend on whether the datasets to be matched are at comparable scales. For example comparing the number of incoming and outgoing arcs of two nodes is relevant only when the data sets have the same scale. At different scales, the cardinality of links in the network changes, making matching networks at different scales meaningless.

2.2 Data matching: The global steps to make a decision

The matching process consists of five steps: pre-processing, candidate selection, criteria combination and decision, and finally evaluation.

Some pre-processing is usually performed on datasets, such as rubber-sheeting to reduce geometric deviation [7, 49]; splitting edges using matched nodes to reduce cardinalities to 1:1 [7,49]; removing nodes having only two arcs connected [46]; or the simplification of the most detailed dataset, if data scales are different [51].

Candidate selection is a first and usually rough analysis to restrict the number of potentially identical features, and thus to reduce computation time. Generally for point and linear features, the selection of candidates is based on geometry using a maximum distance [6, 14, 31, 49] or a buffer [29, 46, 50, 51]. Some authors propose to select candidates for matching using both geometry and semantics [42]; others rely only on semantics—the candidates are matched only if they have similar natures [49]. For polygons, the selection is based on the intersection of features. In other words, all features that intersect each others are candidates [5, 23].

As discussed earlier, many matching criteria can be defined based on similarity measures. Three main approaches to combining criteria to make a decision are defined in the literature: a sequential approach, where criteria are applied one after the other [14]; a combination approach, where criteria are weighted and combined using a sum operator [18]; and a mixed approach. For the latter case, some authors proposed to use a sequential approach to eliminate candidates and to compute, for example, a weighted sum of values issued from each criterion for retained candidates [49, 51]. Other alternatives have considered applying logical rules (e.g., OR) by using thresholds [29]. Generally the "best" candidate in terms of final similarity measure (single or combined) or a matching maximizing the similarity [28] is selected.

The evaluation step is generally an interactive one, with few approaches using automatic evaluation [9, 31, 50]. For a pair of matched features the evaluation can be in terms of "sure" or "doubtful" [9, 31], or in terms of precision and recall [6].

2.3 Discussion: Towards a more generic data matching approach

It is obvious that the geometry, scale, and heterogeneity of data have an important influence on the matching strategy, making the approaches more or less complex. We observe that specific data matching approaches are generally defined according to the geometry of features and/or the scale of datasets to be matched. Nevertheless, most of these processes do not holistically consider both spatial and non-spatial properties. The main consequence is that the data matching is an ad-hoc process, adapted to particular cases.

Despite the emergence of a number of papers exploring the use of semantic information to quantify semantic similarity between concepts [12, 21], only a few approaches are using it to match features, for example, in land use change studies [12]. As far as we know, there are no vector data matching approaches that take into account semantic information in a formal way using concept comparison, and not just string comparison. This type of feature could be useful to minimize aberrant matching links (e.g., matching a valley with a summit, or a motorway with a secondary road), or to avoid eliminating good matching due to the heterogeneities of classifications (e.g., the concepts "summit" and "peak" are semantically close, and might reasonably be matched even if a string comparison considers them different).

We note that data matching approaches are often based on different, clearly defined criteria applied sequentially one after another. This has two advantages: the process is relatively easy to implement and, most importantly, it is easily mastered by users who may control, amongst other things, the effect of parameters. On the other hand there are some cases where analyzing criteria in sequence is not efficient. It can lead either to over-matching or under-matching results, when isolated criteria are either not strict enough or

too strict (e.g., two features may be matched if, simultaneously, they have more or less the same attribute values, are relatively close and have similar shapes).

One key issue when matching relates to imperfection in data, and in particular how to manage imperfect positions, imperfect classifications, errors, or missing data. Imperfections are inherent in vector data. The complexity of the real world leads to different interpretations, and thus errors, imprecision, and missing data when represented in a database. Features that appear in a high quality dataset may not exist in a lower quality dataset. Features may also appear in the lower quality dataset that do not exist in the higher quality dataset. Different scales may lead to different link cardinalities (1:1, 1:n, and n:m). A vector database is always a compromise between cost and quality. Although the issue of imperfections induced by the mapping process has been studied extensively in the field of GIS [20], there is still no widely accepted classification of the different aspects of quality. The different terms being employed include: uncertainty, imprecision, fuzzy, error, and ambiguity.

Reasoning with imperfect knowledge is a topic in artificial intelligence, where imperfections are generally classified in three types: imprecision (regarding the difficulty of clearly and precisely defining a state of reality by a proposal), uncertainty (doubts about the validity of knowledge), and incompleteness (refers to the complete or partial absence of knowledge) [8, 43].

In this context, our goal is to propose general data matching framework that allows computation and fusion of several criteria to match different features, while taking into account data and knowledge imperfection. The approach is a multi-criteria data matching process guided by explicitly formalized knowledge, allowing the fusion of several criteria based on spatial and non-spatial information. Many mathematical theories that manage and formalize imperfections exist in the literature. Among these, *belief theory* is most well-adapted to our needs, and will be the basis of our approach.

3 Some elements of belief theory

The matching process is guided by an explicit model of knowledge. In some cases, knowledge might be missing or unavailable, or it might also be present but unreliable or uncertain. To deal with this, sources of knowledge need to be combined to improve the robustness and quality of the decision making process. This is why numerous different sources have to be combined when matching. Therefore, the challenge is to fuse and model different types of knowledge.

Many methods, such as Bayesian probability, belief theory (also called the Dempster-Shafer model), fuzzy sets, possibility theory, and rough sets are available to manage imperfect knowledge and combine different sources of information. A number of reviews highlighting the advantages and disadvantages exist in the literature. The interested reader is directed to Cohen (1985), Dubois and Prade (1988), and Bouchon-Meunier (1989) [8, 10, 15].

In order to have a generic matching process and to take into account imperfections, a new data matching approach is proposed in this paper based on belief theory. This framework has been introduced by [13] and proven to be an efficient way method to model imperfect knowledge with [43]. Belief theory was chosen here for its major advantages which respond to our needs. Specifically, belief theory:

(1) allows modeling imperfect knowledge (imprecise, uncertain or incomplete);

(2) provides tools to efficiently fuse sources of knowledge to make decisions;
(3) explicitly manages conflicts, i.e., disagreements between sources of knowledge;
(4) allows modeling perfect knowledge and total ignorance; and
(5) allows the use of different metrics to define belief functions.

Belief theory was already used in previous research for classification or map matching in the field of geographic information [11, 32, 40], and for integration of sensor data for decision making purposes [30, 37]. However, as far as we know, there is no data matching approach based on belief theory.

3.1 Framework of belief theory

Belief theory supposes the definition of a finite set of N singleton hypotheses corresponding to the potential solutions of a given problem. This set, generally noted Θ is called a *frame of discernment* and is defined as $\Theta = \{H_1, H_2, \ldots, H_N\}$, where H_i , $i = 1, \ldots, N$ represents a singleton hypothesis.

Let us denote 2^Θ, the power set of Θ, as the set of all possible combinations of the singleton hypotheses belonging to Θ. The power set is defined by:

$$2^\Theta = \{\varnothing, \{H_1\}, \{H_2\}, \{H_1, H_2\}, \ldots, \Theta\} \tag{1}$$

where $\{H_i, H_j\}$ is a subset representing the proposition that the solution of a problem is one of these hypotheses, i.e., either H_i or H_j. A key point of belief theory is the basic belief assignment. A belief assignment is a function that assigns to a proposition $P \in 2^\Theta$ a value, named the mass of belief and denoted $m(P)$, which represents how much a criterion (or "source of information" in the vocabulary of the theory) *believes* in this proposition. As an example, let us consider a process of data matching based on distances between features and a proposition stating that two given features are identical. The closer the two features are, the stronger the criterion believes that they are identical, and the bigger the mass of belief of this proposition is.

A basic belief assignment is a function $m : 2^\Theta \to [0, 1]$ such that:

$$\sum_{P \subset \Theta} m(P) = 1 \tag{2}$$

Belief theory offers tools to combine several sources of information such as the Dempster's rule of combination [13] (see equation 3). Let us consider two sources S_1 and S_2. Each source supports propositions P with a mass of belief, $m_1(P)$ and $m_2(P)$, respectively. We denote by m_{12} the mass of belief resulting from the combination of the two sources by Dempster's rule and that supports the same proposition P.

$$\forall P \in 2^\Theta, m_{12}(P) = \frac{\sum_{P' \cap P'' = P} m_1(P') \times m_2(P'')}{1 - \sum_{P' \cap P'' = \varnothing} m_1(P') \times m_2(P'')}, P' \text{ and } P'' \in 2^\Theta \tag{3}$$

This rule defines how to merge several masses of belief to determine a new mass of belief expressing the combination of beliefs. To make a decision, i.e., to determine which proposition is the solution to the problem, different criteria have to be combined, potentially leading to a conflicting situation (represented by the denominator of equation 3). Belief theory provides different operators to manage this conflict [26, 43, 44].

After the combination of sources, a decision among propositions is made. A simple approach would be to select the candidate C_i having the bigger mass of belief assigned to hypothesis $appC_i$. However, first, there are cases where no candidate is particularly distinguished from the other ones and, second, no overview on all candidates exists, since each candidate is separately analysed. As for fusion, belief theory offers several decision rules developed in literature (e.g., the maximum of plausibility, the pignistic probability). For more details the reader is referred to Smets and Kennes [45].

4 Description of the matching process

In this section, our data matching approach based on belief theory is detailed.

4.1 Data matching approach

Let us consider two vector datasets to be matched. For each feature belonging to one dataset *DataSet1*, matching first consists of looking for potentially identical features in the other dataset *DataSet2*. Then these candidates are analyzed to determine final matching links. Our matching process follows this approach and consists of five steps detailed below. In our case, a source in belief theory framework is a matching criterion.

4.1.1 Step 1: Selection of candidates

The first step consists of defining the frame of discernment. For each feature $F1$ in *DataSet1*, we look for close features in *DataSet2*, according to a distance criterion. These features are the candidates $\{C_i\}_{i=1,\dots,N}$ for matching with $F1$. The frame of discernment is then defined as follows:

$$\Theta_{F1} = \{appC_1, appC_2, \dots, appC_i, \dots, appC_N, NM\} \tag{4}$$

The frame of discernment is the set of hypotheses $appC_i$ expressing that "candidate C_i represents the same real world object than the feature $F1$." Note that in some cases real objects are represented only in one datasets. This case occurs for example when datasets have different actualities (e.g., a new real world object is mapped *DataSet1* but not yet in *DataSet2*) or different mapping specifications. To take into account the case where a feature from *DataSet1* may have not identical feature in *DataSet2*, the hypothesis, *NM*, standing for "feature $F1$ is not matched to any features in *DataSet2*," is added. Thus, the *NM* hypothesis has three main purposes: i) it models cases when real objects are not represented in both datasets; ii) the frame of discernment becomes exhaustive by adding the *NM* hypothesis; and iii) conflicts generated by the fusion of criteria are less important, since the solution necessarily belongs to the frame of discernment.

Example Let us suppose a feature $F1$ in *DataSet1* and two candidates C_1 and C_2. The frame of discernment is then defined as: $\Theta_{F_1} = \{appC_i, appC_2, NM\}$, meaning that $F1$ can be matched with C_1 or C_2 or not-matched at all.

4.1.2 Step 2: Analysis of each candidate independently

To compute the basic belief assignments, a local approach that analyses separately each candidate is used. This approach follows that proposed by Appriou [2], where sources assign masses of belief to only a subset of hypotheses of 2^Θ. In this approach, masses of belief are assigned to hypotheses of all S_i, subsets of 2^Θ defined for each candidate C_i as $S_i = \{appC_i, \neg appC_i, \Theta\}$, where:

(1) $appC_i$ is the hypothesis (singleton) that $F1$ is identical to candidate C_i.
(2) $\neg appC_i = \{appC_1, \dots appC_{i-1}, appC_{i+1}, \dots, appC_N, NM\}$ is the hypothesis (not singleton) that $F1$ is not identical to C_i, i.e., $F1$ is either matched to another candidate, or not matched at all.
(3) $\Theta = \{appC_1, appC_2, \dots, appC_i, \dots, appC_N, NM\}$ is the hypothesis (not singleton) expressing ignorance, i.e., the criterion does not know if C_i is the right candidate or not.

Thanks to the model proposed by [2], both well-known knowledge $m(appC_i)$, uncertainty $m(\neg appC_i)$, and ignorance $m(\Theta)$ are modeled.

Example Following the same example, two subsets are defined for C_1 and C_2:

$$S_1 = \{appC_1, \neg appC_1, \Theta\} \qquad \text{and} \qquad S_2 = \{appC_2, \neg appC_2, \Theta\},$$

where $\neg appC_1 = \{appC_2, NM\}, \neg appC_2 = \{appC_1, NM\}$ and $\Theta = \{appC_1, appC_2, NM\}$.

4.1.3 Step 3: Fusion of criteria

Once masses of belief have been initialized, criteria are combined for each candidate using the Dempster's rule, as defined in equation 3. Thus, for each set of propositions S_i the masses of belief are combined to provide a mass of belief synthesizing the knowledge from the different criteria. At the end of this step, for each candidate, one mass of belief is assigned to each hypothesis $appC_i, \neg appC_i, \Theta$.

Example Following the same example, let us suppose now that two criteria are used to make a decision, criterion 1 and criterion 2. According to the description of step 3, each criterion assigns a mass of belief to each hypothesis. The following masses of belief are obtained:

$m_1(appC_1)$	$m_2(appC_1)$	$m_2(appC_2)$	$m_1(appC_2)$
$m_1(\neg appC_1)$	$m_2(\neg appC_1)$	$m_2(\neg appC_2)$	$m_1(\neg appC_2)$
$m_1(\Theta)$	$m_2(\Theta)$	$m_2(\Theta)$	$m_1(\Theta)$

where $m_1(appC_1)$ and $m_2(appC_1)$ represent the mass of belief respectively assigned by the criterion 1 and 2 to hypothesis $appC_1$.

The computation of the fusion step can be easily illustrated by using a matrix, where the masses of belief assigned by the criteria are arranged into columns (m_1 for criterion 1) and lines (m_2 for criterion 2). Each cell represents the intersection of hypotheses.

	$appC_1$	$\neg appC_1$	Θ
$appC_1$	$appC_1$	ϕ	$appC_1$
$\neg appC_1$	ϕ	$\neg appC_1$	$\neg appC_1$
Θ	$appC_1$	$\neg appC_1$	Θ

The masses of belief related to the candidate C_1 using equation 3 are as follows:

$m_{12}(appC_1) = m_1(appC_1) \times m_2(appC_1) + m_1(appC_1) \times m_2(\Theta) + m_2(appC_1) \times m_1(\Theta)$
$m_{12}(\neg appC_1) = m_1(\neg appC_1) \times m_2(\neg appC_1) + m_1(\neg appC_1) \times m_2(\Theta) + m_2(\neg appC_1) \times m_1(\Theta)$
$m_{12}(\Theta) = m_1(\Theta) \times m_2(\Theta)$
$m_{12}(\phi) = m_1(appC_1) \times m_2(\neg appC_1) + m_1(\neg appC_1) \times m_2(appC_1).$

Note that the same equations are obtained for the candidate C_2.

4.1.4 Step 4: Fusion of candidates

After the fusion of criteria, a set of belief masses for each candidate is calculated. A fourth step that consists of the fusion of masses of belief assigned to each candidate is carried out. Once again, the fusion of candidates is carried out using Dempster's rule, as explained in equation 3. At the end of this step, one combined mass of belief is assigned to each hypothesis of 2^Θ. Note that, if in our approach the *NM* hypothesis is not initialized, it appears during the fusion of criteria, as a result of the contradiction between other hypotheses related on different candidates or different sources.

Example Following the same example, the masses of belief obtained in Step 3 are fused according to the equation 3. The matrix shows the intersections of hypotheses: columns represent hypotheses issued from step 3 for candidate C_1, lines represents hypotheses for candidate C_2. Each cell represents the intersection of hypotheses.

	$appC_1$	$\neg appC_1$	Θ	ϕ
$appC_2$	ϕ	$appC_2$	$appC_2$	ϕ
$\neg appC_2$	$appC_1$	NM	$appC_1, NM$	ϕ
Θ	$appC_1$	$appC_2, NM$	Θ	ϕ
ϕ	ϕ	ϕ	ϕ	ϕ

For example, the mass of belief assigned to $appC_1$ after the combination of the candidates is: $m_{1...3}(appC_1) = m_{12}(appC_1) \times [m_{12}(\neg appC_2) + m_{12}(\Theta)]$.

The *NM* hypothesis appears during this step and it is computed as follows: $m_12(NM) = m_{12}(\neg appC_1) \times m_{12}(\neg appC_2)$.

4.1.5 Step 5: Decision

The final decision is made using the criteria of "maximum of pignistic probability" $P(H)$ (see equation 5) [13]. By choosing the pignistic probability, the decision is made only among the simple hypotheses, i.e., matched to only one candidate ($appC_i$) or not matched (*NM*). However, it is important to mention that this is not only a comparison of masses of beliefs of singleton hypotheses: all propositions that contain a singleton hypothesis are taken into

account in the computation of its pignistic probability, to choose the "best" simple hypothesis.

$$P(H_1) = \sum m(H_m) \frac{|H_1 \cap H_m|}{|H_1|}, \quad H_1 \in 2^{\Theta}, H_1 \subset H_m \tag{5}$$

where $|H_1|$ represents the number of singleton hypotheses contained in H_1.

The hypothesis with the highest mass of belief is chosen in our process. For each chosen hypothesis, a confidence level equal to the difference between the first and the second maximum is computed. Doubtful results (i.e., confidence is less than a threshold, or conflict between criteria is relevant) are highlighted and may be interactively checked.

Example The pignistic probability is computed for the singleton hypothesis: $appC_1$, $appC_2$, and *NM*. The hypothesis with the highest pignistic probability is chosen.

For example, the pignistic probability for $appC_1$ hypothesis is computed as follows:

$$P(appC_1) = m_{1...3}(appC_1) + \frac{|appC_1 \cap \neg appC_2|}{|\neg appC_2|} m_{1...3}(\neg appC_2) \tag{6}$$

4.2 Modeling of data matching criteria

In this section, some typical matching criteria are presented. The belief assignment functions and thresholds illustrated below are typical examples that were tested on actual data. Even though we believe that those functions are quite general, they can be adapted to specific data, and many other criteria may be used.

In the Figures 2 and 3, the first line represents the mass of belief assigned to hypothesis $appC_i$, i.e., "candidate C_i is identical to feature $F1$." The second line shows the mass of belief assigned to hypothesis $\neg appC_i$ meaning "candidate C_i is not identical to feature $F1$." The last line represents the mass of belief assigned to ignorance, $m(\Theta)$.

4.2.1 Knowledge based on geometry

Geometry describes the position, the extension, and implicitly the shape of features, but also captures spatial relationships between features. A criterion based on position is essential (matched features are close features). Nevertheless other criteria may also be considered, such as a criterion comparing orientations of polylines. How to model knowledge for typical position and orientation criteria is shown in Figure 2.

Position criterion This criterion is based on the distance between the positions of two features. This criterion has already been proven to be efficient when based on Euclidean distance between isolated points [33], or based on Hausdorff distance between edges of roads networks [34]. We strongly believe that any position criterion could be used with a similar modeling of masses of belief for other distances, and that thresholds are related to the known accuracy of datasets.

The following example shows how our knowledge about geographic data can be modeled as belief assignment functions for this criterion:

(1) The closer the features are, the more we believe they should be matched together, and the less we believe they should not be matched. This is translated in the globally decreasing (resp. increasing) shape of function for the hypothesis $appC_i$ (resp.

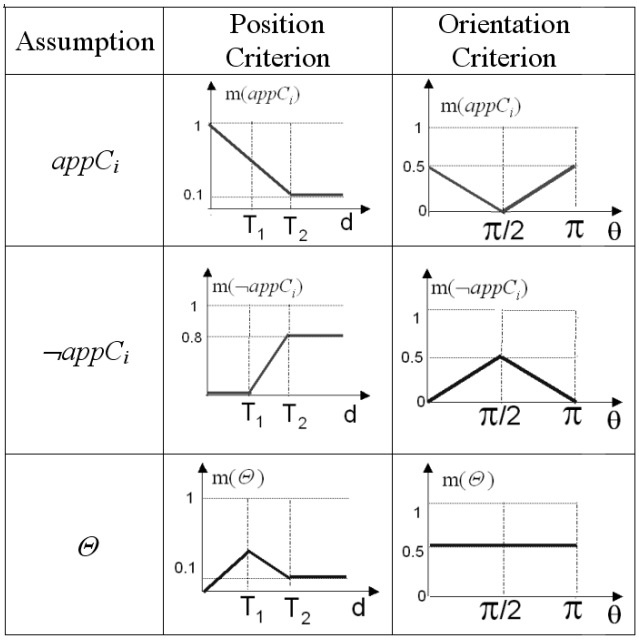


Figure 2: Knowledge modeling for *Position* and *Orientation* criteria.

$\neg appC_i$). This knowledge is "continuous" in the sense that there is no natural distance threshold beyond which the belief should drastically change. More generally, most functions are continuous functions in our approach to avoid thresholds effects.

(2) If the distance between features ($F1$ and candidate C_i) is below the accuracy of the less detailed dataset (threshold T_1), there is no reason to believe that the two features could not be matched, whatever the distance is. This is reflected by the fact that the mass of belief of hypothesis $\neg appC_i$ is zero in this case.

(3) In general, beyond a certain distance (threshold T_2), two features should not be matched together. Therefore, the hypothesis $appC_i$ becomes in this case equally improbable, whatever the distance is, while hypothesis $\neg appC_i$ becomes equally credible. However, this is not always true. Indeed, especially for objects fuzzily defined (e.g., a valley located by a single point), features may be quite distant but identical. Thus the mass assigned to hypothesis $appC_i$ (resp $\neg appC_i$), may not be exactly 0 (resp. 1).

(4) The candidate may neither be close enough to conclude that it is the right identical feature, nor far enough to conclude that is not. This is modeled in the masses of belief function by ignorance. Ignorance is at its maximum for this criterion for intermediate distances.

The great strength of our approach is the ability to explicitly model ignorance. As we restrict the set of considered hypotheses, such as $m(appC_i) + m(\neg appC_i) + m(\Theta) = 1$, defining two of those functions also defines the third. This is convenient for knowledge elicitation, as in certain cases it may be more natural to define certain functions rather than

others: commonsense knowledge may be naturally expressed as "I know that this is it," or "I know this is not it," or "I don't know if this is it."

Orientation criterion An orientation criterion is a typical loose criterion (last column in Figure 2) that consists of comparing local orientations between $F1$ and candidate C_i: colinear features are more likely identical features than perpendicular features. However this is neither a necessary nor a sufficient criterion alone to make a decision. This criterion may be used for linear features. It may be evaluated through on the angular difference θ between the orientations of tangents to $F1$ and to C_i respectively where the point of $F1$ is nearest to C_i, and where the point of C_i is nearest to $F1$. If the angle θ between the two features is close to 0, features are relatively parallel and have the same direction; if the angle θ is closed to π, features are parallel but have opposite directions; finally, if the angle θ is approximately $\pi/2$, features are locally perpendicular. Knowledge about linear data can be translated into functions using the following rules:

(1) Colinear features are more likely identical than perpendicular ones. This explains the global shape of functions for hypotheses $appC_i$ and $\neg appC_i$.
(2) However, the orientation criterion may not be sufficient to make a decision, even if it gives some clues for the final decision. Especially when datasets have different levels of detail, many identical features may not have the same orientation. Consequently, ignorance has a significant weight. In the example in Figure 2, the mass of belief assigned to ignorance has the same value whatever the value of the angle is. This is a typical way of easily giving more or less importance to some criteria.

4.2.2 Knowledge based on semantic information

To compare the different semantics of features, a semantic distance, d_S is needed. Semantic distance may be computed using Wu and Palmer distance [4]. In our experiments this distance is computed using a geographic ontology, obtained by automatic extraction from textual specifications of the two datasets [1]. From a semantic point of view, if the semantic distance is close to 0, then two features have similar natures (e.g., peak and summit). If the semantic distance is close to 1, these features are very different (e.g., river and summit).

Semantic criteria are typical "necessary but not sufficient" criteria. Two features may only be matched if they have close natures; but all features having the same nature are not necessarily matched. Figure 3 illustrates how to model such a semantic criterion:

(1) The "not sufficient" part is modeled by sharing the mass of belief between the hypothesis $appC_i$ and ignorance when features have close natures.
(2) The "necessary" part is modeled by assigning a significant mass of belief to hypothesis $\neg appC_i$, meaning that we strongly believe that features are not identical when they have very different natures. More drastically, if the semantic distance is beyond a threshold T_S, the two features will definitely not be matched. This knowledge is represented by a constant function for all hypotheses. The mass of belief assigned to hypothesis $\neg appC_i$ is significant. However, if we want to deal with data with possible classification errors, the masses of belief assigned to the hypotheses $appC_i$ and ignorance may be close to 0, but not exactly 0.

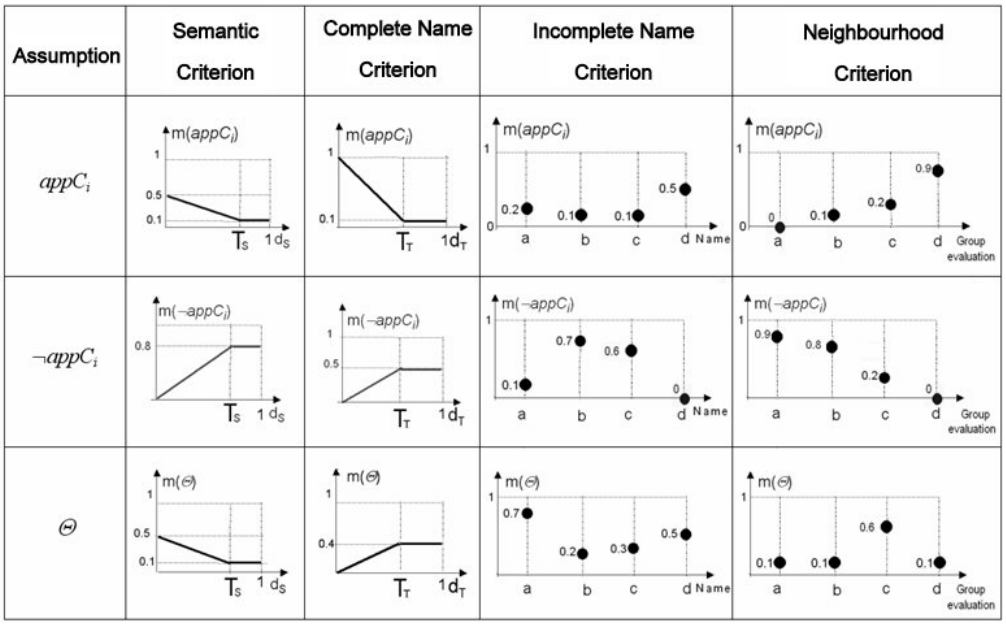


Figure 3: Knowledge modeling for *Semantic*, *Complete Name*, *Incomplete Name*, and *Neighborhood Criteria*.

4.2.3 Knowledge based on names

Some attributes representing the name of the features (e.g., toponym or postcode) or the name of groups of features (e.g., road number) may be considered as pseudo-identifiers. However in some cases, these names may be only sparsely filled in the database (e.g., names of features may be only filled in for important objects). Moreover, in some cases vernacular names can co-exist, such as with places. To model that, let us describe two criteria to compare names. The *Complete Name Criterion* is especially adapted to name attributes filled in, but possibly imprecise, such as toponyms. The *Incomplete Name Criterion* is adapted to incomplete but precise data such us road name). Both criteria are based on string comparison using Levenshtein distance, noted d_T [27].

Complete name criterion As illustrated in the third column in Figure 3, the name criterion is significant when two names are similar or strongly similar, since that name may be considered as an identifier.

(1) If distance is zero, (i.e., both features have exactly the same name), we believe that the two features are identical, and the mass of belief assigned to hypothesis $appC_i$ is equal to 1. Otherwise, the mass of belief decreases proportionally with distance. Note that an underlying assumption of no homonyms is made here. If duplicate names exist, two situations are identified: i) a conflict situation may appear at the combination step if other characteristics of pairs to compare are different; or ii) a matching link is defined otherwise. The first situation is highlighted in our process since the conflict is used here to manage difficult situations. The second situation implies that an object

to match, F_1 is matched with two candidates. Being a part of the global evaluation of results, the last case is easy to detect. When names are not similar, this criterion is less significant leaving the possibility to others criteria to decide. Therefore if the distance d_T is greater than the threshold T_T (e.g., 30% of letters are different) the hypotheses $\neg appC_i$ and ignorance are credible, the hypothesis $appC_i$ being improbable. By making ignorance a significant hypothesis, the modeling of the *Complete Name Criterion* is cautious, i.e., if we are sure, we make a decision on a hypothesis, otherwise we are impartial.

(2) Cases of ambiguity are modeled by assigning to hypothesis $appC_i$ a mass of belief different from 0 in any case (for cases like when two names represent the same object in the real world, one having the official name whereas the other has a non-official name, or when names are expressed in different dialects).

Incomplete Name Criterion The fourth column in Figure 3 shows the mass of belief assignments for the *Incomplete Name Criterion*. For this criterion, datasets reflecting close points of view (i.e., features may be simultaneously thought of as "important" or "not important") are considered. Hence, we believe that names of identical features should be simultaneously filled in or not. Another hypothesis would lead to a different definition of belief. Four cases are considered when feature $F1$ and candidate C_i are compared:

(1) Case A: The attribute is not filled in both datasets. In this case, a decision cannot be made, assigning a major mass of belief to ignorance. The complement of ignorance is divided into hypotheses $appC_i$ and $\neg appC_i$. More credibility may be given to $appC_i$ than to $\neg appC_i$, as the fact that the name is not filled in both datasets may reflect that the two features have a similar level of importance.

(2) Case B: Only one feature has a name value filled in. It is not impossible that the two features are identical, so we assign to the hypothesis $\neg appC_i$ a relevant mass of belief, with a low ignorance.

(3) Case C: Features have different names. In this case we believe that C_i is not an identical feature, and a significant mass of belief is assigned to $\neg appC_i$, with a low ignorance. To manage cases in which vernacular features co-exist, the hypothesis $appC_i$ is not completely rejected. Instead a low, but not null, mass of belief being assigned to it.

(4) Case D: Features have the same names. It is highly probable that the features are identical. However, there are cases when an object of the real world is represented by many feature in datasets (e.g., roads are represented by line-segments). Cases when $F1$ has candidates with the same name frequently appear. Thus, we assign the same mass of belief to $appC_i$ hypothesis and to ignorance.

4.2.4 Knowledge based on the analysis of neighborhood

To take into consideration the holistic nature of matching (i.e., matching a feature depends on the matching of its neighbors), a *Neighborhood Criterion* (see the last column in Figure 3) may also be defined. To do that, the process should actually be slightly more complex than described in Section 4.1. In a first iteration, the matching process is performed using some criteria, then the results are used to initialize the masses of belief of the neighborhood criterion as explained below. In a second iteration, the final matching is performed with

the same criteria plus the neighborhood criterion. For complex cases, the process can even be reiterated again.

Let us consider a feature $F1$ belonging to dataset *DataSet2* (the most detailed dataset) and a candidate C_i belonging to dataset *DataSet1* (the less detailed dataset). According to Section 4.1, 0 or n links are generated by the first step, i.e., each feature of *DataSet1* may be matched with 0 or n features in *DataSet2*. Then, for each candidate, C_i, belonging to *DataSet1*, their n identical features defined in the first iteration are grouped into connected groups. If only one group has been identified, this group is evaluated as being sure that is to say that all features belonging to the group, including $F1$, are considered to be well-matched. Otherwise, if several groups are found neighbors of C_i are analyzed to see how they are matched in the first step, and especially if corresponding features of the neighbors are connected to the groups. Four cases are distinguished (see Figure 4 for examples of these cases and last column of Figure 3 for knowledge modeling):

(1) Case A: None of the neighbors of C_i is matched with a neighbor of $F1$ in the first iteration. We believe that C_i and $F1$ are not identical.
(2) Case B: One of the neighbors of C_i is matched with a neighbor of $F1$. In this case, we weakly believe that C_i is identical to $F1$.
(3) Case C: Several but not all the neighbors of C_i are matched with neighbors of $F1$. Hence, we strongly believe that C_i is identical to $F1$.
(4) Case D: All the neighbors of C_i are matched with neighbors of $F1$. Thus, we most strongly believe that C_i is identical to $F1$.

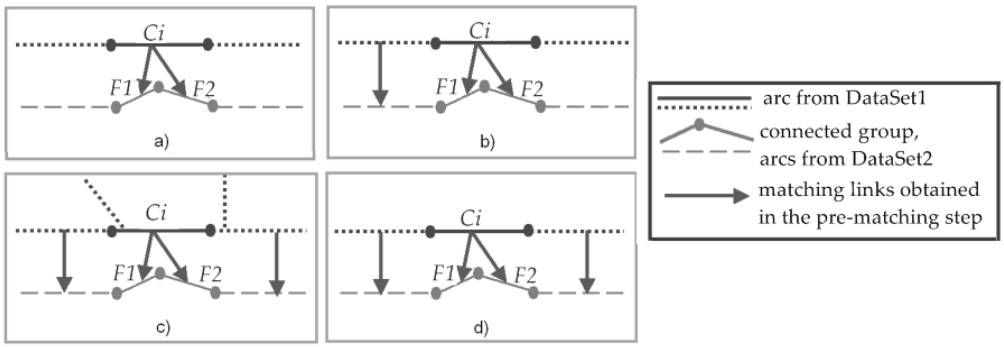


Figure 4: Different cases of matching results for networks.

5 Experimentation and evaluation

This section describes some experimentation on two types of data: isolated points and road networks. Implementation was carried out using the COGIT Laboratory platform GeOxygene[1]. GeOxygene is an open platform based on Java that implements OGC/ISO specifications for the development and deployment of GIS applications.

[1]GeOxygene, http://oxygene-project.sourceforge.net

We note that our matching approach can also be applied on polygons by using the same criteria for attribute or semantic information and by defining new criteria for geometry. For example, criteria based on surface distance [5] or shape comparisons could be proposed [3]. With regards to this criteria, the more two shapes are overlapping or have similar shapes, the more chances there are to be matched.

5.1 Isolated points: Presentation of the case study and typical results

Experiments on isolated points have been performed on relief data extracted from two datasets of IGN-France, the French National Mapping Agency (NMA): BDTOPO© and BDCARTO©, and with the following characteristics. These datasets have different levels of detail, with BDTOPO being more detailed than the BDCARTO dataset. Features representing relief such as mountains, summits, and valleys are vague, firstly by definition (e.g., boundary between a valley and a mountain is not perfectly defined) and secondly because the classification of objects may be confusing (e.g., the difference between a "summit" and a "peak" may be hard to define). In addition, identical features may have different names due to linguistics or due to the use of both official names and vernacular names for the same object. The semantics of features do not have the same level of detail in the databases. For example in the BDCARTO there are concepts, which are grouped together: e.g., "summit, crest, hill," while in BDTOPO these same concepts are well separated. Using only the position criterion is not always efficient since the identical feature is not always the closest one. In the same way, using only the *Toponym Criterion* may lead semantic inconsistencies. Therefore our matching approach is based on the fusion of the three criteria (*Position, Toponym*, and *Semantic*) seems well adapted to solve this matching problem.

Quantitative results of the matching process are described below. Before, Figure 5 illustrates the importance of semantic information in the matching process.

Figure 5a shows that the process did not succeed to match identical features ("l'escarpu ou pic de louesque" and "l'escarpu") when using only the name and position criteria. This is because those features are not very close and their names are slightly different. Instead, identical features are matched when the Semantic Criterion is added (see Figure 5b). This result is partially due to the fact that features have the same nature, and thus the semantic criterion "confirms" the two other "unsure" criteria.

5.2 Experiments on road networks: Typical results

This section describes some experiments matching road networks from two datasets, BDCARTO from IGN and MultiNet© from TomTom. These datasets have different scales, producers, and purposes. They are also highly heterogeneous. Our test area covers approximately 760km^2 in urban and rural areas.

The first difference between the two datasets is scale. MultiNet is more detailed than BDCARTO. However, there are also objects such as pathways and rugged ways that are represented in the less detailed database BDCARTO but not in MultiNet.

Differences in modeling and representation also exist. Each database has a specific representation of the real world according to its purpose. BDCARTO is built by the French NMA. It is used to make maps at 1:100,000 or 1:250,000 scale, and for geographical analyses at regional and departmental levels. It has an accuracy ranging from one to several decameters. MultiNet is built by TomTom, a private company. Its accuracy goes from five to

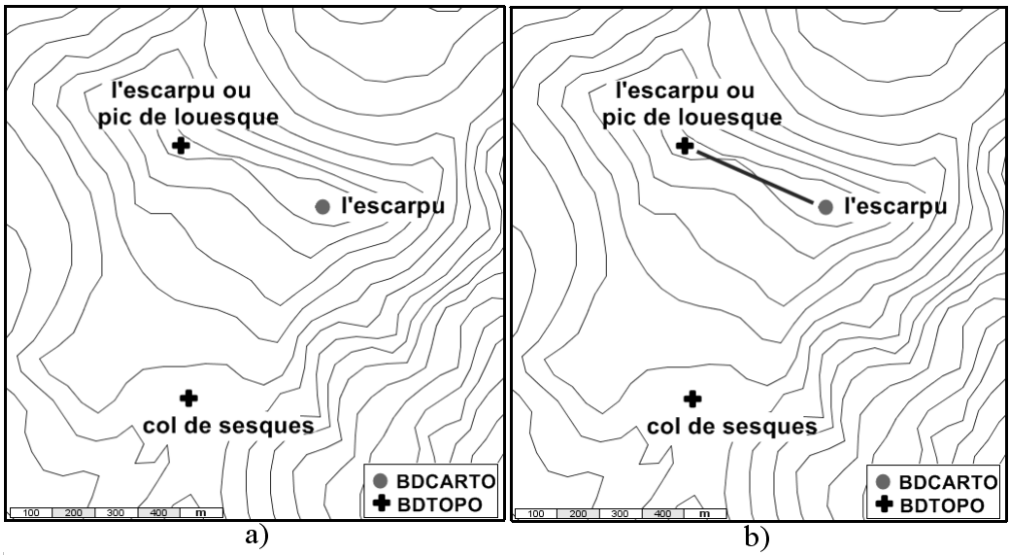


Figure 5: Isolated points matching results using a) *Position* and *Toponym* criteria and b) *Position, Toponym* and *Semantic* criteria.

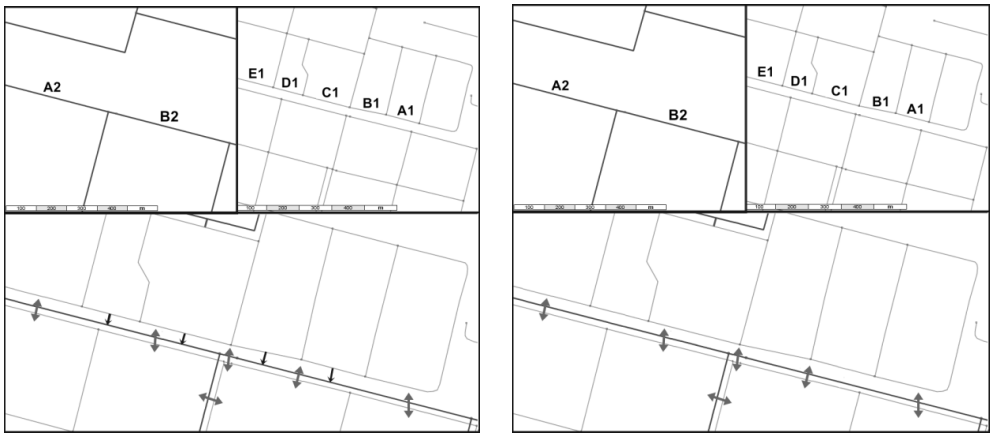
twelve meters in many detailed sections, such as city street networks. The latter database focuses particularly on road and street network description for navigation applications. Thus, highways are modeled by a single polyline representing its axis in BDCARTO, and by two parallel polylines representing the axis of each single carriageway in MultiNet. Another important difference between the datasets concerns the roundabouts. In MultiNet these are represented by a set of edges, whereas in BDCARTO they are generally represented by a single node.

Some results obtained by the matching process described in Section 4.2 are illustrated. For each feature belonging to MultiNet, the most detailed dataset, the algorithm looks for candidates in BDCARTO before choosing the "best" one according to our approach based on belief theory. In the next figures BDCARTO is displayed on the upper left side, while MultiNet is displayed on the upper right side, and both datasets as well as matching links are displayed on the bottom.

Recall that for road networks the matching process is carried out in 2 steps. In a first step, the matching process is performed using *Position, Semantic, Name*, and *Orientation* criteria as described in Section 4.2. The distances are computed between edges of roads (e.g., for each edge to match a distance is computed between it and edge candidates). The *Position Criterion* is based on the Hausdorff distance between edges of roads networks. The *Orientation Criterion* is based on the orientation of an edge candidate compared with the edge to match. The *Semantic* and *Name* criteria are respectively based on the edge types and names. Then, in a second step, the results of the first step are used to initialize the masses of belief of the *Neighborhood Criterion*. The final matching is performed with all five criteria: *Position*, *Semantic*, *Name*, *Orientation*, and *Neighborhood*.

Figure 6 reveals the utility of using many criteria and notably the *Neighborhood Criterion*. Spatial context has a particular importance in the matching process of networks. Thus if

the context is not taken into account i.e., *Neighborhood Criterion* is not used, over-matched matching results occur, especially in urban areas. Thus, segments C1, D1 and A1, B1 are wrongly matched with A2 and B2 segments respectively. This result is due to the datasets having different scales; the segments representing streets are very close and have the same orientations. Therefore the *Semantic* and *Name* criteria are important. Unfortunately our two datasets have heterogeneous classifications in urban areas, and name attributes are not always filled in. As a result we are faced with imprecise and insufficient semantic and name criteria. As a result the matching process without the *Neighborhood Criterion* is not able to distinguish between identical and non-identical features, but when a more holistic analysis is made by adding the *Neighborhood Criterion*, over-matched results are eliminated. Figure 6b shows that the segments E1, D1, C1, B1, A1 are not matched any more, i.e., no identical feature are found in the less detailed dataset. Similarly, Figure 7 illustrates the utility of the *Name Criterion*.



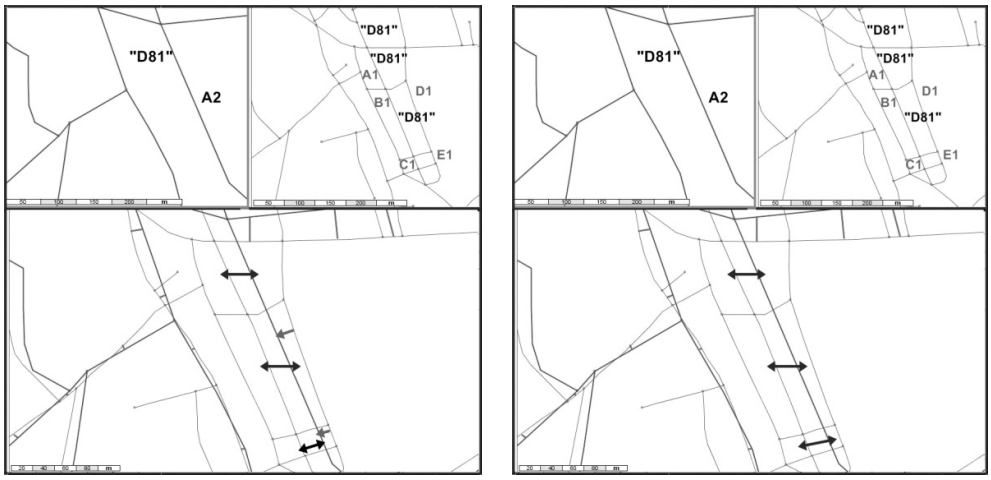
(a) (b)

Figure 6: Matching results with (a) and without (b) the *Neighborhood Criterion*. In both subfigures, the less detailed dataset (BDCARTO) is represented in the upper left corner, and the most detailed dataset (MultiNet) in the upper right corner. Both dataset are overlapped in the bottom. Single arrows represent wrong-matched links and double arrows represent correct links.

In Figure 7a, the results of data matching algorithm without using the *Name Criterion* are shown. The edge A2 representing the road named "D81" is matched with edges A1, B1, D1, and E1. Links between (A2, D1) and (A2, E1) are wrong, because D1 and E1 are not edges belonging to the road "D81." This mistake is rectified when the *Name Criterion* is added to the matching (Figure 7b).

5.3 Qualitative evaluation of matching results

The automated matching results are compared with an interactive matching to evaluate the approach. Evaluation for isolated points is made in relation to the number of features. Five datasets (1,232 features) for five French counties are tested. For roads networks, evaluation is carried out in comparison with the total length of networks (12,725km for a 760km^2 area).



(a) (b)

Figure 7: Matching results with (a) and without (b) using the *Name Criterion*. In both subfigures the less detailed dataset (BDCARTO) is represented in the upper left corner, and the most detailed dataset (MultiNet) in the upper right corner. Both dataset are overlapped in the bottom. Single arrows represent wrong-matched links and double arrows represent correct links.

The evaluation is made on behalf of the most detailed dataset and regarding to the ground truth for isolated points and an interactive matching for road networks.

Two well-known indicators used in information retrieval domains to measure the performance of algorithms were applied to qualify our results. These are precision—the percentage of the matching links defined by the matching algorithm that are correct compared to an interactive matching, and recall—the percentage of the correct matching links that are actually discovered by the matching algorithm. The recall and precision were compared with two approaches: Beeri et al. [5], which defines for isolated points, and only based on distances between features (see Table 1); and Mustiere and Devogele [31], dedicated for road networks at different levels of detail, and based on geometric and topological analysis (see Table 2).

			Number of features	Precision	Recall
Isolated points	Ground truth	Matched	1,131	100%	100%
		Non-matched	101	100%	100%
	Our approach	Matched	1,126	99%	99%
		Non-matched	92	91%	95%
		Conflict	14	-	-
	Beeri et al. [5]	Matched	1,063	94%	93%
		Non-matched	55	54%	59%

Table 1: Qualitative evaluation of results for isolated points.

For isolated points, among 1,232 features belonging to BDTOPO, 1,131 features should be matched and 101 features should be not-matched (i.e., have no identical features in BDCARTO). Good precision and recall are obtained in our approach, especially for matched features (99%). Conflicts (i.e., no decision is taken) were noted for 14 features. Relatively good precision (91%) and recall (95%) are obtained when Beeri et al.'s [5] approach is used. However, both measures are weak for non-matched features. The Beeri et al. [5] approach does not use name or semantic information to match, and thus has a tendency to match the closest features together.

		Number of edges in MultiNet	% of the length of the MultiNet	Precision	Recall
Our approach	Matched	6093	59%	96%	95%
	Non-matched	6632	41%	95%	94%
Mustière and Devogele [28]	Matched	5304	54%	98%	90%
	Non-matched	7421	46%	85%	98%

Table 2: Qualitative evaluation of results for road networks.

Due to the complexity of networks, precision and recall are lower (96% resp. 95%) for road networks. Comparing with the Mustière and Devogele [31] approach, which is rather pessimistic, our matching approach is optimistic with the precision being lower than the recall. Regarding non-matched features, the results for precision and recall are reversed: our approach has better precision (95% against 85%) and lower recall (94% against 98%) than the Mustière and Devogele [31] approach. Errors (4%) are mainly due to complex roundabouts, which unfortunately are not efficiently managed by our process because in BDCARTO, roundabouts are represented by a single node, whereas in MultiNet they are represented by a set of edges. To improve this, one solution could be to add a new criterion specialized on roundabouts previously detected and to match edges to nodes.

In our opinion the results were slightly improved regarding the other matching approaches thanks to both the use of different criteria and the way the knowledge were formalized, especially ignorance. In this direction, some sensitivity analyses were made in order to measure the importance of using different criteria and parameters used in the process. Concerning the criteria, our approach was tested using different configurations, such as *Distance Criterion* and *Semantic Criterion* or *Distance Criterion* and *Name Criterion*, or the three criteria together. Sensitivity tests have shown that the matching results using our approach are most effective where more criteria are used.

Concerning the sensitivity of parameters, only a few tests were carried out. The analysis consisted of varying the threshold for one criterion while the others were left unchanged. The first results have shown that sensitivity is more important for the *Distance Criterion* and less important for those based on semantic and names. Concerning the *Distance Criterion* we observed that the sensitivity is more important for thresholds corresponding to small distances. This result makes sense since datasets have different scales and different precision. Certainly, further tests should be carried out. One possibility would be to study the impact of all parameters at the same time, and with respect to knowledge about datasets such as precision, accuracy etc.

Conclusion

In this paper a data matching approach based on knowledge fusion using belief theory is proposed. We showed that the matching process becomes more efficient thanks to the fusion of different knowledge and more generally, thanks to an explicit formalization of that fused knowledge. The addition of knowledge, even if incomplete, leads to an improvement of the matching process (e.g., name attribute). In conclusion, we consider that our data matching approach has the following advantages.

First, it can be used on different data, i.e., the approach can be adapted according to type of data (e.g., points, polylines, polygons), to scale (the same or different scales), and to different layers (e.g., relief, buildings, road networks, hydrographical networks).

Second, the matching approach can be seen such as an adaptive and expandable process. It is possible to add as many criteria as are needed without modifying the approach defined in Section 4.1.

Finally, imperfection is explicitly modeled in our approach. Imperfection is represented by the belief masses assigned to a proposition (i.e., the union of singleton hypothesis). Uncertainty is explicitly modeled by the partial belief masses (i.e., $appC_i, \neg appC_i, \Theta$). Finally incompleteness is managed by assigning the unit mass of belief to ignorance (i.e., the hypothesis defining the frame of discernment). In our opinion, inclusion of ignorance is the greatest strength of our approach because it allows the use of criteria that are relevant for some cases, and not relevant for others. Another advantage of ignorance is in ergonomics: the user can fix two of the three functions, and deduce the third one. In some cases, the easiest way is to quantify the "no matched" and "no opinion," as in the case of the orientation test. Quantifying the "no matched" means "the less features that are colinear, the less matching is possible." Quantifying "no opinion" expresses the low conviction of this criterion. In other cases, it is easier to quantify the "matched" and "not matched," such as for the position criterion where the first means "two very close features are probably identical," and the second means "two distant features in any case more than the precision of datasets, are probably not matched."

The assignment of masses of belief is an essential step in our approach. As we have noted in Section 4, curves defining the belief assignments vary from criterion to criterion. They also have different weights in the process. This is due to the fact that each criterion relies on different knowledge that is more or less perfect. This flexibility is another key advantage of our approach as it allows users to precisely model different knowledge, However it may also be thought of as a drawback, as tuning the process may become fastidious. In our experiments, thresholds depend on data specifications and especially on the precision of a dataset and are thus directly set up. Moreover, in our opinion, the definition of precise thresholds is not so important for two main reason: many criteria are combined, and the curves are relatively smooth. As such only approximate thresholds are necessary. In addition, we believe that even if thresholds may be adapted to special cases, the global shapes of functions are quite general, and should be reused in many matching cases.

For future work, there are several possible solutions that could make the matching process more generic and easier for end users. One solution is to develop a method for optimizing parameters that is a compromise between matching results quality and number. A second solution is to identify thresholds and weights by data mining. Finally, a third solution is to carry out a qualitative study for each criterion in order to evaluate it.

On final point to discuss is the cardinality of matching links. Our approach is "one-way," with each edge of one dataset (the most detailed) matched to one edge in the other dataset (the less detailed). It may happen that several edges of the detailed network are matched to the same edge of the other network. However, if the networks have similar scales this should not happen. To improve this aspect, post-processing could be constructed. First, pignistic probabilities for the two features matched with the same candidate could be compared, and then the feature with the highest probability chosen. Second, a new criterion allowing taking into account spatial context could be introduced. For example, Samal et al. [38] propose a matching approach that analyses the local systematic shifts between matched features. This idea could be introduced as a new criterion, in a similar way to the neighborhood criterion.

References

[1] ABADIE, N., MECHOUCHE, A., AND MUSTIÈRE, S. Owl-based formalisation of geographic databases specifications. In *Proc. 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW'10)* (Lisbon, Portugal, 10 2010), pp. 11–15.

[2] APPRIOU, A. Probabilités et incertitudes en fusion de données multi-senseurs. *Revue Scientifique de Technique de la Défense 11* (1991), 27–40.

[3] ARKIN, E. M., CHEW, L. P., HUTTENLOCHER, D. P., KEDEM, K., AND MITCHELL, J. S. B. An efficiently computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*, 3 (1991), 209–216. doi:10.1109/34.75509.

[4] BALLATORE, A., BERTOLOTTO, M., AND WILSON, D. The semantic similarity ensemble. *Journal of Spatial Information Science 7* (2013), 27–44. doi:10.5311/josis.2013.7.128.

[5] BEERI, C., KANZA, Y., SAFRA, E., AND SAGIV, Y. Object fusion in geographic information systems. In *Proc. 30th International Conference on Very Large Databases* (Toronto, Canada, 2004), vol. 30, VLDB Endowment, pp. 816–827. doi:10.1016/b978-012088469-8/50072-3.

[6] BEL HADJ ALI, A., AND VAUGLIN, F. Geometric matching of polygons in GIS and assessment of geometrical quality of polygons. In *Proc. International Symposium on Spatial Data Quality (ISSDQ'99)* (Polytechnic University, Hong Kong, 1999), W. Shi, M. Goodchild, and P. Fisher, Eds.

[7] BLASBY, D., DAVIS, M., KIM, D., AND RAMSEY, P. GIS conflation using open source tools. OpenJump White Paper, 2004.

[8] BOUCHON-MEUNIER, B. On the management of uncertainty. *Encyclopedia of Computer Science and Technology 20*, 5 (1989), 327–337.

[9] CLODOVEU, D., AND FONSECA, F. Assessing the certainty of locations produced by an address geocoding system. *GeoInformatica 11*, 1 (2007), 103–129. doi:10.1007/s10707-006-0015-7.

[10] COHEN, P. R. *Heuristic Reasoning About Uncertainty: An Artificial Intelligence Approach*. Pitman Publishing, Inc., Marshfield, MA, USA, 1985.

[11] COMBER, A., FISHER, P., AND BROWN, A. Uncertainty, vagueness and indiscernibility: The impact of spatial scale in relation to the landscape elements. In *Proc. International Symposium of Spatial Data Quality (ISSDQ'07)* (Enschede, Holland, 2007).

[12] COMBER, A., FISHER, P., AND WADSWORTH, R. Assessment of a semantic statistical approach to detecting land cover change using inconsistent data sets. *Photogrammetric Engineering and Remote Sensing 70*, 8 (2004), 931–938. doi:10.14358/pers.70.8.931.

[13] DEMPSTER, A. Upper and lower probabilities induced by multivalued mapping. *Annals of Mathematical Statistics 38*, 2 (1967), 325–339. doi:10.1214/aoms/1177698950.

[14] DEVOGELE, T., PARENT, C., AND SPACCAPIETRA, S. On spatial database integration. *International Journal of Geographical Information Science 12*, 4 (1998), 335–352. doi:10.1080/136588198241824.

[15] DUBOIS, D., AND PRADE, H. Representation and combination of uncertainty with belief functions and possibility measures. *Computer Intelligence 4*, 2 (1988), 244–264. doi:10.1111/j.1467-8640.1988.tb00279.x.

[16] DUCKHAM, M., AND WORBOYS, M. An algebraic approach to automated geospatial information fusion. *International Journal of Geographical Information Science 19*, 5 (2005), 537–557. doi:10.1080/13658810500032339.

[17] DUMENIEU, B. Automatic reconstruction of spatio-temporal data from historical maps. In *Proc. Workshop on Integrating 4D, GIS, and Cultural Heritage* (Leuven, Belgium, 2013).

[18] DUNKARS, M. Matching of datasets. In *Proc. 9th Scandinavian Research Conference on Geographical Information Science* (Espoo, Finland, 2003), pp. 67–78.

[19] EGENHOFER, M., CLEMENTINI, E., AND DI FELICE, P. Evaluating inconsistencies among multiple representations. In *Proc. 6th International Symposium on Spatial Data Handling (ISSDQ'94)* (Edinburgh, UK, 1994), pp. 901–920.

[20] FISHER, P. Models of uncertainty in spatial data. In *Geographical Information Systems*, vol. 1. Wiley, 2003, pp. 191–203.

[21] FOODY, G. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing 70*, 5 (May 2004), 627–633. doi:10.14358/pers.70.5.627.

[22] GIRRES, J.-F., AND TOUYA, G. Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS 14*, 4 (8 2010), 435–460. doi:10.1111/j.1467-9671.2010.01203.x.

[23] GOMBOSIA, M., ZALIKA, B., AND KRIVOGRADA, S. Comparing two sets of polygons. *International Journal of Geographical Information Science 17*, 5 (11 2003), 431–443. doi:10.1080/1365881031000072627.

[24] KILPELÄINEN, T. Maintenance of multiple representation databases of topographic data. *The Cartographic Journal 37*, 2 (12 2000), 101–107. doi:10.1179/0008704.37.2.p101.

[25] KOUKOLETSOS, T., HAKLAY, M., AND ELLUL, C. Assessing data completeness of VGI through an automated matching procedure for linear data. *Transactions in GIS 16*, 4 (2012), 477–498. doi:10.1111/j.1467-9671.2012.01304.x.

[26] LEFEVRE, E., COLOT, O., VANNOORENBERGHE, P., AND D., D. B. A generic framework for resolving the conflict in the combination of belief structures. In *Proc. 3rd International Conference on Information Fusion (Fusion'2000)* (Paris, France, 2000), vol. 1, IEEE, pp. MOD4/11–MOD4/18.

[27] LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR 4*, 163 (1965), 845–848.

[28] LI, L., AND M.F., G. An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion 2*, 4 (7 2011), 309–328. doi:10.1080/19479832.2011.577458.

[29] LÜSCHER, P., BURGHARDT, D., AND WEIBEL, R. Matching road data of scales with an order of magnitude difference. In *Proc. XXIII International Cartographic Conference* (Moscou, Russia, 2007), International Cartographic Association. doi:10.5167/uzh-77809.

[30] MAUSSANG, F., ROMBAUT, M., CHANUSSOT, J., HÉTET, A., AND AMATE, M. Fusion of local statistical parameters for buried underwater mine detection in sonar imaging. *EURASIP Journal on Advances in Signal Processing 2008* (2008), 876092. doi:10.1155/2008/876092.

[31] MUSTIÈRE, S., AND DEVOGELE, T. Matching networks with different levels of detail. *GeoInformatica 12*, 4 (10 2008), 435–453. doi:10.1007/s10707-007-0040-1.

[32] NASSREDDINE, D., ABDALLAH, F., AND DENOEUX, T. Map matching algorithm using belief function theory. In *Proc. 11th International Conference on Information Fusion (Fusion'08)* (Cologne, Germany, 2008), IEEE, pp. 995–1002.

[33] OLTEANU, A.-M. A multi-criteria fusion approach for geographical data matching. In *Proc. 5th International Symposium on Spatial Data Quality* (Enschede, Netherlands, 2007).

[34] OLTEANU-RAIMOND, A.-M., AND MUSTIÈRE, S. Data matching—A matter of belief. In *Headway in Spatial Data Handling*, A. Ruas and C. Gold, Eds., Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, Montpellier, France, 2008, pp. 501–519.

[35] OSM-WIKI. OpenStreetMap Project Wiki. http://wiki.openstreetmap.org/wiki/Main_Page, 2014.

[36] PLUMEJEAUD, C., MATHIAN, H., GENSEL, J., AND GRASLAND, C. Spatio-temporal analysis of territorial changes from a multi-scale perspective. *International Journal of Geographical Information Systems 25*, 11 (2011), 1597–1612. doi:10.1080/13658816.2010.534658.

[37] POROSEVA, S., LETSCHERT, J., AND YOUSUFF HUSSAINI, M. Application of evidence theory to quantify uncertainty in hurricane/typhoon track forecasts. *Meteorology and Atmospheric Physics 97*, 1–4 (2007), 149–169. doi:10.1007/s00703-006-0249-9.

[38] RODRIGUEZ, A., AND EGENHOFER, M. Comparing geospatial entity classes: An asymmetric and context similarity measure. *International Journal of Geographical Information Systems 18*, 3 (2004), 229–256. doi:10.1080/13658810310001629592.

[39] RODRIGUEZ, M., AND EGENHOFER, M. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering 15*, 2 (2003), 442–456. doi:10.1109/TKDE.2003.1185844.

[40] ROYÈRE, C., GRUYER, D., AND CHERFAUI, V. Data association with belief theory. In *Proc. 3rd International Conference on Information Fusion (Fusion'00)* (Paris, France, 2000).

[41] SAFRA, E., KANZA, Y., SAGIV, Y., AND DOYTSHER, Y. Efficient integration of road maps. In *Proc. 14th ACM International Symposium on Advances in Geographic Information Systems (ACMGIS'06)* (Arlington, United State, 2006), ACM Press, pp. 59–66. doi:10.1145/1183471.1183483.

[42] SAMAL, A., SETH, S., AND CUETO, K. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Systems 18*, 4 (2004), 459–489. doi:10.1080/13658810410001658076.

[43] SHAFER, G. A. *Mathematical Theory of Evidence*. Princeton University Press, 1967.

[44] SMETS, P. Imperfect information: Imprecision and uncertainty. In *Uncertainty Management in Information Systems*. Springer, 1997, pp. 225–254.

[45] SMETS, P., AND KENNES, R. The transferable belief model. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, R. Yager and L. Liu, Eds., vol. 219 of *Studies in Fuzziness and Soft Computing*. Springer Berlin Heidelberg, 2008, pp. 693–736.

[46] STIGMAR, H. Matching route data and topographic data in a real-time environment. In *Proc. 10th ScanGIS (ScanGIS'05)* (Stockholm, Sweden, 2005).

[47] SUI, H., LI, D., AND GONG, J. Automatic feature-level change detection for road network. In *Proc. 20th ISPRS Congress* (Istanbul, Turkey, 2004).

[48] UITERMARK, H., OOSTEROM, P., VAN MARS, N., AND MOLENAAR, M. Propagating updates: Finding corresponding objects in a multi-source environment. In *Proc. 8th International Symposium on Spatial Data Handling (SDH'98)* (Vancouver, Canada, 1998), pp. 580–591.

[49] VOLTZ, S. An iterative approach for matching multiple representations of street data. In *Proc. ISPRS Workshop, Multiple Representation and Interoperability of Spatial Data* (Hanovre, Germany, 2006), pp. 101–110.

[50] WALTER, V., AND FRITSCH, D. Matching spatial data sets: A statistical approach. *International Journal of Geographical Information Science 13*, 5 (1999), 445–473. doi:10.1080/136588199241157.

[51] ZHANG, M., SHI, W., AND MENG, L. A generic matching algorithm for line networks of different resolutions. In *Proc. ICA Workshop on Generalisation and Multiple Representations* (La Corogne, Spain, 2005).