Dempster-Shafer theory for object matching under data imperfection constraints: Application to wastewater networks' line matching

Yassine Belghaddar, Ahlame Begdouri, Nanée Chahinian, Abderrahmane Seriai, Omar Et-targuy, Carole Delenne

Dempster-Shafer theory for object matching under data imperfection constraints: Application to wastewater networks' line matching

Yassine Belghaddar[1,2,3,4], Ahlame Begdouri[1], Nanée Chahinian[2], Abderrahmane Seriai[3], Omar Et-targuy[1], Carole Delenne[2,4]

1. LSIA , Univ. Sidi Mohamed Ben Abdellah, Fez, Morocco
2. HSM, Univ. Montpellier, CNRS, IRD, Montpellier, France
3. Berger-Levrault, Labège, France
4. Inria Lemon, CRISAM – Inria, Sophia Antipolis Méditerranée, France

Abstract

The goal of object matching is to identify objects representing the same real entity across multiple spatial datasets. This involves comparing and linking data from different sources using similarity measures, with the final matching decision made by combining these measures. Object matching is especially valuable for creating accurate and complete spatial datasets for underground networks, where data often come from various sources and may have imperfections like imprecision or incompleteness. The Dempster-Shafer (DS) theory, which uses mass functions to model data imperfections, is considered the best method for combining imperfect data. However, previous DS-based approaches produced highly conflicting results when many potential candidates for an object existed. In this work, we present an improved DS-based line matching approach for wastewater networks. Our key contributions include introducing candidate ranking, bidirectional measure combination, and mixed models to convert similarity measures into masses. We validated our approach through

experiments on both synthetic and real-world datasets. The results demonstrate that our contributions significantly reduce conflict and improve the accuracy and correctness of the matching decision.
Kewords: Object Matching; Dempster-Shafer theory; Similarity measure; Data integration; Data imperfections

1 Introduction

Wastewater underground systems are mandatory in our modern society. They are used daily to collect and transport effluents to a treatment plant. The choice of burying the networks underground is mainly to minimize accidental or intentional damage to the infrastructure and to ensure the safety of the citizens. Beyond the safety concerns, with a limited surface space, using underground space helps cities meet the increasing demands while remaining compact [1]. Various examples highlight the intensive use of the underground space. For example, France counts more than 2.7 million kilometers of buried and underwater infrastructure [2] while in the USA, the drinking water infrastructure system alone is composed of 3.5 million kilometers of pipes, most of which are underground [3].

To ensure continuous services and proper functioning of treatment plants, wastewater networks often undergo repair and expansion operations. Regardless of the type of network (water, gas, electricity, etc.), being buried makes the inspection and detection of damage difficult. Indeed, the impacts of the incidents undeniably cause budget overruns and affect our communities via traffic congestion, service interruption, or compromising the safety of the individuals, especially the contractors working in the field [4]. On average, 12 deaths and 60 injuries are caused annually by gas pipeline incidents in the USA [5]. In the UK, utilities' street works have a direct impact on traffic congestion [6]. In France, every year, 100,000 network damages occur during intervention [7].

Given that the lifespan of the network is several decades, managers and operators may change over time. Reparation and extension operations may not be properly tracked and reported. Thus, data recording the position of the network's objects and the history of interventions may be collected from different sources with disparate contents and formats, such as digital maps, images, or textual documents.

Precise data about buried utilities are essential for the different actors

working on the networks for distinct purposes. For excavators, it is to avoid damaging underground infrastructures or interrupting traffic while executing interventions [6]. For the entity in charge of the network, it is to reduce management costs by planning and efficiently anticipating the replacement of the network's elements, since urgent and unexpected operation costs are far higher than anticipated ones [8]. For public health and environmental specialists, it is to protect water resources against pollution by using hydraulic simulation software to detect anomalies in existing networks, or to size future ones.

In addition to the heterogeneity of collected data, imperfections may appear, namely inconsistency (such as out-of-use pipelines which still appear on the maps), missing attribute values for some objects, or uncertain and sometimes contradictory values. All of these aspects generate complexity in terms of data control and analysis, and highlight the need for merging multi-source data to obtain more precise and complete digital maps of wastewater networks.

In the literature, combining multiple spatial data sources is referred to as data integration or data conflation. Data integration is defined as the process of unifying existing data sources into a single framework, where the output is a unified description of the sources' schemas, allowing access to the input databases' instances [9]. Data conflation is defined as the process of creating a new dataset based on multiple datasets that cover the same spatial area [10]. Whether the goal is to create a schema to query input instances or to create a new dataset from available sources, object matching is identified as the most difficult and challenging step in both data integration [11] and conflation [12].

Multiple approaches have been proposed for object matching. They all use similarity measures to discriminate between the possible candidates to match. Therefore, what differentiates them is how they define, use and combine these measures. The most widely-used approaches range from traditional methods that rely on thresholds and weighted measures [11], probabilistic approaches [10], to optimization-based techniques that frame the problem as an optimization task and apply off the shelf algorithms to determine the best solution [13]. Although these approaches, particularly optimisation ones, can produce efficient matching results, they do not take into consideration the imperfect aspect of input data, which has a direct impact on the final matching decision.

Several theoretical frameworks for representing and processing uncer-

tain/imprecise information exist. Probability theory has long been considered as the only way to measure and take into account uncertainty. The new theories of uncertainty, developed in the second half of the 20th century, including fuzzy sets, possibilities, and belief functions through the classic Dempster-Shafer (DS) theory, make it possible to better take into account vague, uncertain, imprecise, and paradoxical information collected from different sources. Indeed, while the fuzzy set theory is intended to represent imprecision or vagueness, the Dempster-Shafer theory is considered more powerful, due to its ability to represent both data uncertainty and imprecision.

In the last three decades, significant effort has been deployed to address the problem of fusing or combining data collected from different sources. Data fusion techniques *combine data from multiple sensors, and related information from associated databases, to achieve improved accuracies and more specific inferences than could be achieved by the use of a single sensor alone* [14]. Data fusion techniques have been used in several applications. First, the focus was on the military domain, such as target tracking [15], then it was extended to multiple fields, such as urban water modelling [16], Artificial Intelligence (AI) [17], and many others.

The DS theory is widely used to perform data fusion by combining evidence collected from different sources. It has been used in various applications such as pedestrian detection [18], data annotation [19], natural disaster management [20] and optical coordinate metrology [21]. To generate high-quality, large-scale data essential for AI systems, the authors in [17] applied DS theory to integrate information from multiple sources. As for object matching, the DS theory has been adopted as a general framework where the mass functions model the imperfections related to the similarity measures. Indeed, the step of mass initialization plays a key role in the process of matching and directly influences the final decision. In this context, two main existing models, proposed in [22] and [23], have been used to initialize the masses in [24], [25], [26], resulting in highly conflicting output values when the number of candidates is large. This is particularly relevant in wastewater networks, where the number of potential candidates for each object can exceed 10. In such cases, a matching decision can not be made, and the previous approaches should not be directly applied. The aim of our work is to propose an enhanced DS-based process to achieve object matching of wastewater spatial data, taking into account data imperfections.

The main contributions of our work are: i) a deep analysis of the na-

ture of data imperfections related to the domain of wastewater networks, ii) proposition of a generic approach for line matching in the context of spatial data, iii) significantly reducing the conflict by introducing candidates' ranking, bidirectional measure combination and mixed models to transform the similarity measures to masses in the combination step, and iv) experimental validation of the effectiveness of the proposed approach.

This paper is organized as follows: Section 2 presents data imperfections in the wastewater networks as well as related works in the field of object matching in general, and under imperfection constraints in particular. Section 3 presents background on DS theory and its use in object matching. Our approach is introduced in Section 4. Experiments and results are presented and discussed in Section 5. Section 6 concludes this work.

2 Research background

2.1 Wastewater data representation

A wastewater network is composed of different objects. It is represented by a graph composed of nodes and edges. Nodes represent manholes, equipment, repairs, etc., and the edges represent pipes. Each of the nodes and edges has a set of properties in the form of attributes such as diameters of the pipes, types of materials, and positions.

Wastewater data are usually stored, visualised, and exploited through Geographic Information Systems (GIS) software such as QGIS and ArcGIS. Shapefile and GeoJSON are among the most popular formats to store geospatial data, particularly wastewater network data. Currently, hydraulic simulation software such as SWMM can manipulate wastewater data in their GIS format for purposes as diverse as the sizing of new networks, the simulation of treatment plant hydraulics, flood prevention, and so on.

Although the applications are various, the digital representation of the data remains almost identical and can be divided into two categories:

- Spatial data indicating the position and the geometry (point, line, polygon) of the network objects (junctions, manholes, and pipes). Figure 1 shows an example of a wastewater network stored in a Shapefile format and visualized in QGIS software.

- Attributes characterizing the network objects, such as pipes' diameters

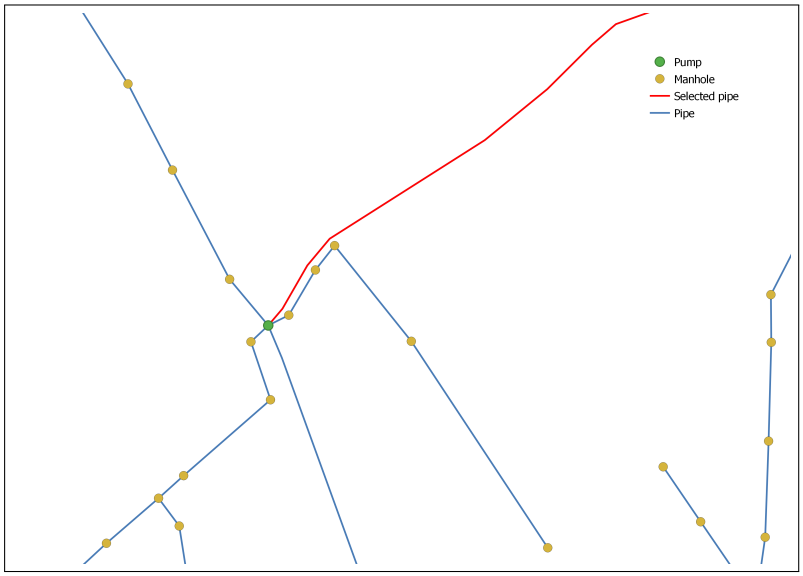


Figure 1: Example of the spatial representation of wastewater network in QGIS.

or materials. They are usually stored in alphanumeric attribute tables, where each record is associated with a network object. Table 1 shows the attribute table associated with the Shapefile of Figure 1.

OBJECTID	COMMUNE	DATE_POSE	DIMENSIONS	MATERIAU	ECOUL
100	JACOU	2999-12-31	200	PVC	GRAVITAIRE
101	JACOU	2999-12-31	200	PVC	GRAVITAIRE
102	GRABELS	2999-12-31	200	AMIANTE CIM.	GRAVITAIRE
104	GRABELS	2999-12-31	300	FONTE	GRAVITAIRE
106	CLAPIERS	2999-12-31	150	AMIANTE CIM.	GRAVITAIRE
111	CLAPIERS	2999-12-31	160	PVC	GRAVITAIRE
112	GRABELS	2999-12-31	200	AMIANTE CIM.	GRAVITAIRE

Table 1: Example of the spatial representation of wastewater network.

2.2 Data imperfections

Despite the standardization of geospatial data representation, imperfections still appear among wastewater datasets. There is no rigid definition of data

imperfections. Depending on the domain of study or the application, imperfection could represent different forms of issues related to the data: vagueness, uncertainty, ambiguity, confusion, etc. In the literature, various classifications of imperfect data have been proposed [**?**], [27]. For wastewater networks, three types of imperfections are considered:

- Incompleteness: indicates missing information.
- Imprecision: the doubt about multiple possible values.
- Uncertainty: related to the veracity of an assertion, the information can be precise and complete, but false.

One of the most important and frequent cases of data imperfection in wastewater network data concerns incompleteness, more precisely, missing nodes in a branch of the network, which can result from several scenarios. In some junctions (the part that connects two pipes is called a junction), equipment is installed for different purposes. For instance, a manhole is built to inspect the pipes. The junctions with no equipment (manhole, valve, etc.) or ramification of the network are often partially recorded by data providers. As a first scenario, data could have been lost or not recorded when the network was first installed. As a second scenario, some users are more interested in the global structure of the network rather than the details about the networks' components. Finally, in an attempt to complete the data, signal-based inspection methods, which are the most popular techniques to collect missing data of buried assets, can be used. When no equipment is installed on a junction, it is difficult for these techniques to identify the transition between two pipes. Consequently, nodes representing the junctions can be missing from the wastewater network's graph, as illustrated in Figure 2.

The issue of missing junctions is not specific to wastewater networks and can be noticed in multiple networks, such as gas [28]. In addition, the junctions of some geographical objects, such as rivers [29] or roads [30], may not be explicitly represented in spatial databases. To cope with missing nodes and make networks more structurally comparable when conducting spatial data matching, researchers often rely on the concept of strokes, as in [31].

Indeed, underground networks are all characterized by their specific topology since the main branches are connected to each other and ramified into several sub-branches. These hierarchical relationships are important to handle the issue of missing nodes. A stroke is defined as a "good continuity"

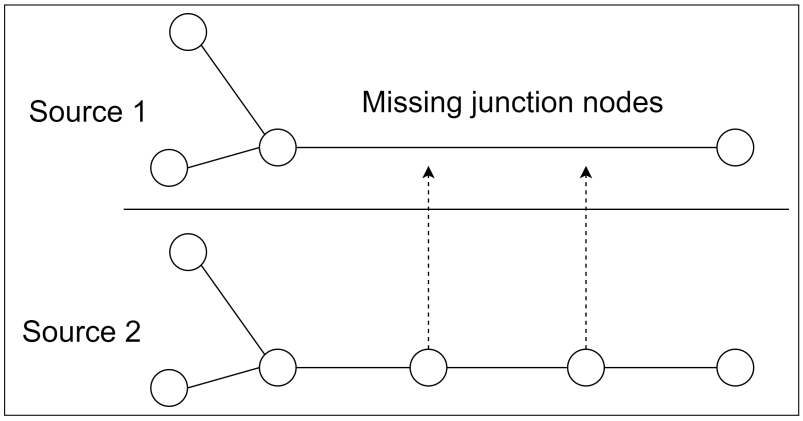
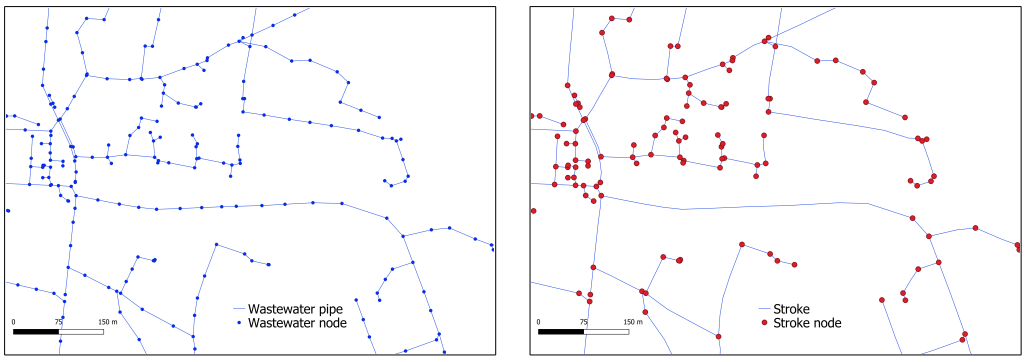


Figure 2: Example of missing junction nodes.

between edges that do not exceed a degree of deviation [28], [32]. Different criteria can be added to the angle of deviation to define the "good continuity", such as the best deviation angle when multiple choices are available [33] or using semantic and toponymic information [32]. Therefore, when nodes are available in one dataset and absent in the others, using the stroke concept will capture the overall structure of the networks and help achieve matching regardless of missing nodes. Figure 3 shows an example of strokes created from a graph.

2.3 Object matching

Several works on object matching have been performed in numerous domains such as buildings [34], [35], road networks [11], [30], hydrographic networks [31], underground networks [28], and many others. Since matching operations are motivated by different goals, the proposed methods are diverse and cannot be listed exhaustively. To navigate through the available ones in the literature, we can differentiate between unidirectional matching, from one dataset (generally called reference) to another dataset (called target) [36], and bidirectional matching, where the datasets are used simultaneously as reference and target [35]. We can also categorize the available solutions according to: i) the supported objects' representation: nodes (e.g. [34]), lines (e.g. [28]) or polygons (e.g. [36]), some propositions support all three representations as in [10]; ii) the supported representation scale: some studies



(a) Before stroke generation. (b) After stroke generation.

Figure 3: Example of strokes created from a graph.

address datasets with identical spatial scales [30] and others with different scales [9], [35], [36]; iii) the supported cardinalities: in general we distinguish between the methods that accept only (1:1) and (1,0) cardinalities, i.e. one object can match at most with one object in the other dataset, and the methods where the (1:N),(M:N) and (M:1) cardinalities are also supported such as in [35] and [36]; iv) the number of datasets that can be matched: methods can support either 2 (e.g. [36]) or N> 2 (e.g. [37]) datasets.

Object matching is conducted following different approaches and methods which, generally, share two main components: i) identifying similarity measures between the network objects, and ii) defining a matching process/steps that uses these measures to match the objects.

2.3.1 Distances vs. similarity measures

To assess whether two or more spatial objects from distinct databases represent the same real entity, comparison criteria, referred to as similarity measures, are required and form an essential part of any matching approach. Similarity measures and distances are complementary concepts. A distance is what is measured directly between two objects, such as the Euclidean distance. Generally, a distance is not directly used by matching methods, instead, it is converted to a similarity measure by normalizing its value [37]. The main objective of this transformation is to have a common scale for the different distances and to simplify computation. Hence, two objects are more likely to correspond when the similarity between them is high. In the

literature, similarity measures in object matching concern the three essential characteristics of the network and the objects to be matched: the geometric, topological, and attribute aspects.

- **Geometric similarity measures** are based on geometric criteria, such as the position, the form, or the angle:
  - The Euclidean distance is the most popular position criterion for node matching [34].
  - The Hausdorff and the Fréchet distances both allow to measure the distance between multi-point geometries. They are the main geometric distances for line matching [11], [30], [38]. Given two point sets A and B, the Hausdorff distance is computed as the maximum distance selected from the minimum distances between each point a of A to point b of B. Unlike the Hausdorff distance, the Fréchet distance considers the ordering of the points along the shapes, which makes it suitable for comparing curves, but harder to compute [39]. A small outlying point from one point set leads to an important increase in the Hausdorff distance. To minimize the impact of outlying geometry, an extension of the Hausdorff distance based on the median Hausdorff distance was proposed in [40].
  - To compare the form of the geometry, the length is used for lines as in [11], and the area for polygons as in [37].
  - The angle is usually used to measure the orientation of lines, as in [25]. It is computed by the differences in angles or directions of two or more objects compared to a fixed axis [10], [37].

  In addition to these popular geometric distances, others can be developed and used, such as the perimeter of a triangulation as proposed in [36].

- **Topological measures** capture the relationships between an object and its neighbours. The node degree measure, i.e., the number of edges connected to a node, is one of the common topological measures. The node degree is used to check whether two or more geometries have the same neighbourhood structure. It is often used for road matching [11], [12], [41]. Several other topology-based measures were proposed.

For example, in [37], assuming that the geographical context of the objects to match is invariant, a proximity graph is drawn for each object, and a comparison between these graphs is established to achieve building matching. In [28], a spatial scene composed of the neighbours of the objects is used as a topological measure to match underground networks.

- **Attribute metrics** compare the attribute values of the objects to be matched, such as addresses or names. The Levenshtein distance is one of the common distances in this category [31], [37]. It is defined as the minimum number of insertions, deletions, and substitutions needed to transform one character string to another. In addition to the value of the attributes, their semantics is often used when it is necessary. For example, in [25], a conceptual similarity between object types is analysed to minimize aberrant matching, such as matching a valley with a summit. In [42], a semantic-similarity step, based on word embedding, is used to achieve instance matching and is applied on real-world benchmark datasets from the Ontology Alignment Evaluation Initiative (OAEI) for restaurants, health, and drugs.

2.3.2 Object matching process

Historically, the US Census Bureau was one of the first organizations to initiate matching objects in order to achieve map conflation of separate digital maps of metropolitan areas [41]. Since then, several studies have been published, and significant progress has been made.

The same real-world spatial data can be stored and represented in multiple datasets. Updating and propagating changes manually, between multi-representation datasets, often leads to inconsistency and is time and money-consuming. To address this issue, researchers in [35] proposed an object matching process that identifies identical features to automatically propagate changes. Their method was applied to building maps from China, where each building is represented by a polygon. Distance, area, direction, and length were used to assess the similarity between the polygons.

To analyse and to understand the relation between the topography of the French territory and the evolution of its population distribution, authors in [31] compared old and new hydrographic maps. Given significant discrepancies between networks, varying levels of detail, and the inaccuracy of old

data, a line matching process was developed to identify homologous parts. To make the networks more structurally comparable, lines were transformed into strokes. Frechet distance, toponymic distance, and orientation were considered as criteria to judge whether two strokes are similar.

A similar approach was developed in [28] to identify the pipes of a gas network. The objective was to capture the overall structure of the network and ease data sharing between multiple providers.

When it comes to matching entities represented by points, authors in [34] indicate that the location of objects is the only property that is always available. Using only the Euclidean distance between the objects as a similarity measure, the method is applied to find corresponding hotels and tourist attractions, represented by points, from different maps.

Matching methods are usually influenced by the dataset at hand and the purpose of the matching, hence the diversity of the propositions. However, most of the propositions have three main steps in common: candidates' selection, similarity measures' combination, and decision.

- **Candidates' selection**: consists in selecting the closest objects, since corresponding ones are usually spatially close, and therefore, reducing the number of potential candidates for the matching. This is achieved by using a buffer (threshold on Euclidean distance), and/or a set of filters (thresholds applied to any other measures) usually defined by an expert. For each object in the reference dataset, the candidates are then defined as the set of objects from the target dataset that fall within the buffer/filters. Buffer selection is widely used in object matching [11], [30], [31], [36], [37]. As for filters, the candidates are generally those having a measure value lower than a fixed threshold [30], [34], [35], [37].

- **Similarity measures' combination**: for each candidate of a reference object, the set of used measures must be combined to allow the matching decision. The most popular combination scheme is a weighted average [10], [35], [37], [43]. The weights associated to each measure indicate the capacity of the measure to discriminate between the candidates. These weights are usually defined by domain experts, but can also be learned from samples of data, as in [35], where the weights are learned using an Artificial Neural Network (ANN). Another common combination scheme consists in decreasing the filters' thresholds iteratively until only one candidate at most remains [43]. Other more

advanced methods to compute the combination can be found in the literature, such as defining an average value based on a compatibility value in [12], or defining a score named the global priority value in [31].

- **Decision**: the result of the matching relies directly on the decision method, as two decision methods can provide two different outcomes, despite using the same similarity measures. At this step of the matching process, for each object that has one or more possible candidates left, the goal is to use computed scores in the previous steps to decide which candidate(s) is likely to be the corresponding object(s). This includes cases where the object has no corresponding entity(ies). The most direct method to decide consists in choosing the 'maximal' or the 'minimal' value from the combination step [31], [43]. Optimization methods are popular in this field. In [37], the authors reformulated the problem in graph theoretical terms, precisely the well-known problem of finding the maximal cliques of a graph. The nodes of the graph represent the objects to match, an edge between two nodes represents a possible corresponding object, and the values of the edges correspond to a weighted similarity between the objects. To find the maximal clicks, the authors used shelf methods from graph theory. In [30], after defining a set of similarity measures, the matching is transformed to an optimization problem, where the goal is to maximise an objective function based on a set of defined constraints.

2.4 Object matching under imperfection constraints

Object matching is challenging not only because of the differences in terms of resolution, schemas, temporalities, and representations between the sources to be matched, but especially because of data imperfections such as incompleteness, distortion, and imprecision. Managing imperfections comes down to answering two questions: i) How to model the imperfections? ii) How to reflect these imperfections on the matching results?

To perform object matching, objects from different datasets are compared using different criteria such as the position or the form. The matching is decided by combining them. Depending on the imperfection of the data at hand, the contribution of each criterion may vary. For example, when the datasets suffer from important distortion, the contribution of a distance-based criterion is less important than the one based on the topology.

As described in the previous paragraph, although available optimization approaches can produce efficient matching results in many cases, they were not designed to handle data imperfections. However, several theoretical frameworks for representing and processing uncertain/imprecise information and making decisions in a context of uncertainty exist. Probability theory constitutes the classic and historical framework for representing and processing uncertain information. It has long been considered as the only way to measure and take into account uncertainty. The new theories of uncertainty, developed in the second half of the 20th century, including fuzzy sets, possibilities, and belief functions through the classic Dempster-Shafer theory, make it possible to better take into account vague, uncertain, imprecise, and paradoxical information collected from different sources. Indeed, while the fuzzy set theory is intended to represent imprecision or vagueness, the Dempster-Shafer theory is considered more powerful, due to its ability to represent both data uncertainty and imprecision.

In the framework of the DS theory, degrees of confidence (or masses of belief) are associated with the validity of information based on the lower and upper limits of a family of probability distributions. These masses no longer apply to singletons only, but also to subsets. In addition, several ready-to-use operators are available to conduct the combination of multisource data.

All these arguments make the DS theory very suitable for application in object matching under imperfection constraints. In the following paragraph, we present the basic concepts of this theory as well as some examples of its application in the field of object matching.

3 Dempster-Shafer's theory

3.1 Main conecpts

Combining information within the framework of DS theory requires the definition of a common fusion space or a frame of discernment $\Omega = \{\mathrm{H}_1, \mathrm{H}_2 \ldots \mathrm{H}_\mathrm{N}\}$, that denotes the set of discrete, mutually exclusive and exhaustive possible N hypotheses or events. The likelihood associated to a subset of Ω is defined by the mass function $m(.)$:

$$m : 2^\Omega \mapsto [0, 1] \text{ with } \sum_{A \subseteq \Omega} m(A) = 1 \tag{1}$$

$m(A)$ represents the degree of belief attributed to the proposition A which could not, given the available knowledge, be assigned to a subset more specific than A. A is called focal element when $m(A) > 0$. In addition, since A is a subset of Ω, $m(A)$ represents the imprecision when the cardinality of A is greater than one. The mass value of a subset $A = \{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ is not equally distributed between the singletons $\mathrm{H}_1, \mathrm{H}_2$ and H_3, contrary to the probability theory. Hence, the value associated to $m(\Omega)$ represents the ignorance and not an equiprobability.

The reasoning mechanisms are grouped into two levels: credal and pignistic. At the credal level, two functions are defined to measure the total belief delivered by a source of evidence:

- The belief function Bel(.) defines the minimum likelihood associated to A as:

$$\mathrm{Bel}(\emptyset) = 0,\ \mathrm{Bel}(\Omega) = 1 \text{ and } \mathrm{Bel}(A) = \sum_{B \subseteq A} m(B) \tag{2}$$

  It represents the mass of all the information provided by a source that supports A.

- The plausibility function Pl(.) defines the maximum likelihood associated to a subset A of Ω as:

$$\mathrm{Pl}(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{3}$$

  It represents the mass of all the information that does not contradict A, or the mass that could be transferred to A if the information available was less uncertain.

At the pignistic level, the pignistic probability related to a hypothesis $\mathrm{H_i}$ is defined in [44] as follows:

$$\mathrm{BetP}(\mathrm{H_i}) = \sum_{A \subseteq \Omega, \mathrm{H_i} \in A} \frac{m(A)}{|A|} \tag{4}$$

where $|A|$ is the cardinal of A. It represents a measure of subjective probability that respects the property:

$$\forall A \subseteq \Omega, \mathrm{Bel(A)} \leq \mathrm{BetP}(A) \leq \mathrm{Pl}(A) \tag{5}$$

Once the information collected from different sources is represented by mass functions, DS theory offers several operators to combine them. Here we describe three popular operators:

- Conjunctive rule of combination: let $m_1(.)$ and $m_2(.)$ be two mass functions representing information about two independent and reliable sources. The conjunctive operator is defined by:

$$(m_1 \cap m_2)(A) = \sum_{B \cap C = A} m_1(B) \times m_2(C), \quad A \subseteq \Omega \tag{6}$$

  The conjunctive rule of combination is generally used when the sources are independent and reliable.

- Normalized conjunctive rule of combination:

$$m_1 \oplus m_2 = \begin{cases} \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) \times m_2(C) & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \tag{7}$$

  Where K is the mass associated to the conflict between the two masses m_1 and m_2, defined by:

$$K = m(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{8}$$

  The normalized rule is generally used to distribute the conflict on the hypotheses according to their masses.

- Disjunctive rule of combination: despite not generating conflict, this rule is less precise than the conjunctive one, since the resulting combination sets are larger than the initial sources. This rule is defined by:

$$m_1 \cup m_2(A) = \sum_{B \cup C = A} m_1(B) \times m_2(C), \quad A \subseteq \Omega \tag{9}$$

  The disjunctive rule is used when the sources are unreliable. It enlarges the focal elements and therefore loses specificity, which can render the final decision difficult.

To transfer information from one frame of discernment to another, one can use the refinement operation R, which associates to each hypothesis

of a frame of discernment $\Omega^1 = \{A_1^1, A_2^1 \dots A_k^1\}$ a subset $R(\mathrm{H_i})$ of another frame of discernment $\Omega^2 = \{A_1^2, A_2^2 \dots A_p^2\}$ such as $\{R(A_1^1), \dots R(A_k^1)\}$ form a partition of Ω^2, with:

$$\forall A \subseteq \Omega^1 \; m^2(R(A)) = m^1(A) \tag{10}$$

In the context of DS theory, when information about the reliability of a source is available, a discounting operation can be applied to the mass function to integrate this information. For a degree of reliability $r \in [0, 1]$ a discounting rate of value $\alpha = 1 - r$, modify $m(.)$ as follows:

$$m^\alpha(A) = (1 - \alpha)m(A) \tag{11a}$$

$$m^\alpha(\Omega) = (1 - \alpha)m(\Omega) + \alpha \tag{11b}$$

Finally, the DS theory offers several rules of decision, such as selecting the hypothesis with the highest belief, plausibility, or pignistic probability value.

3.2 Dempster-Shafer theory in object matching

In the context of object matching, the DS theory has been previously applied in [24], [25], [26]. The mass functions are used as a tool to model the imperfections related to the similarity measures, such as the imprecision of objects' positions. In order to identify corresponding objects, a multi-criteria data matching approach guided by a formal representation and fusion of sources using the DS theory is proposed in [25]. After extracting candidates for each reference object from the comparison dataset; assuming that $\mathrm{H_i}$ is the hypothesis i that the reference object corresponds to a candidate $n \in \{1, 2 \dots N\}$ where N is the number of objects in the comparison dataset, and that $\Omega = \{\mathrm{H}_1, \mathrm{H}_2 \dots \mathrm{H}_N\}$ is the set of all possible hypotheses, a set of distances are computed and transformed into mass functions using the following model:

$$m = \begin{cases} m(\{\mathrm{H_i}\}) = v_1 \\ m(\neg\{\mathrm{H_i}\}) = v_2 \\ m(\Omega) = v_3 \end{cases} \tag{12}$$

where v_1, v_2 and v_3 are initialized based on subjective knowledge about the considered distances. For example, the closer two features are in terms of

Hausdorff distance, the more important the mass v_1 should be, with $v_1 + v_2 + v_3 = 1$. Inspired by the work of A. Appriou in [45], a two-level fusion step is carried out to decide which couples to match:

- For each potential couple, the initialised mass functions based on the similarity distances are combined using the normalised combination rule where $\Omega_i = \{\mathrm{H_i}, \neg \mathrm{H_i}\}$ is the local frame of discernment related to the hypothesis i. This combination results in affecting new mass values to $\mathrm{H_i}$ and $\neg \mathrm{H_i}$ based on the information given by all of the elementary similarity measures between the reference object and its candidate objects.

- For each reference object and based on the refinement operation, the mass functions of all the potential couples resulting from the first step are combined based on the global frame of discernment $\Omega = \{\mathrm{H_1}, \mathrm{H_2} \ldots \mathrm{H_N}\}$. This second combination allows to affect new mass values to all of the possible hypotheses. Then, a decision is made based on the highest pignistic probability

To accurately estimate the position of a moving vehicle on a road, the authors in [26] combine three different mass functions. Two of them are defined directly based on formulas derived from hypotheses about the domain knowledge. The third mass function is intended to combine GPS, sensor, and digital map data and is computed based on a similarity criterion. Assuming that $\mathrm{H_i}$ is the hypothesis of being on a road i, that $\Omega = \{\mathrm{H_1}, \mathrm{H_2} \ldots \mathrm{H_N}\}$ is the set of the hypotheses related to all N possible roads and that $\{\neg \mathrm{H_i}\}$ is the complement of $\{\mathrm{H_i}\}$ in Ω, a likelihood value L_i of the vehicle being on road $i \in \{1, 2 \ldots N\}$ is determined based on the intersection between the current vehicle position and the road i. The likelihood is transformed into mass functions using the following model:

$$m_i = \begin{cases} m_i(\{\mathrm{H_i}\}) = 0 \\ m_i(\neg\{\mathrm{H_i}\}) = r_i(1 - \mathrm{L_i}) \\ m_i(\Omega) = 1 - r_i(1 - \mathrm{L_i}) \end{cases} \tag{13}$$

where r_i is a discounting coefficient associated to a road i. A three-level fusion step is then carried out to decide which road to match:

- The mass values related to each road i are initialized with $\Omega_i = \{\mathrm{H_i}, \neg \mathrm{H_i}\}$ as a local frame of discernment and then, transferred to the global frame of discernment Ω.

- The mass functions of the candidates of Ω are combined using the normalized combination rule. This results in affecting masses values to all the hypotheses of Ω, which defines the 3rd mass function.
- The three mass functions are combined, using also the normalized conjunctive rule. The decision is taken based on the highest pignistic probability.

Similarly to the previous work, the authors in [24] combined multiple criteria using DS theory to match spatial objects (Points of Interest). In this study, four similarity measures were considered: spatial similarity based on the Euclidean distance, name similarity based on the Levenshtein distance, address similarity, and the category of the point of interest. Contrary to the previous proposition, where the mass functions initialization was related to the local frame of discernment $\Omega_i = \{\mathrm{H_i}, \neg \mathrm{H_i}\}$ sets the value of $\mathrm{H_i}$ to 0, in this study the model applied to transform the likelihood into mass values is the following:

$$m_i = \begin{cases} m_i(\{\mathrm{H_i}\}) = r_i(\mathrm{L_i}) \\ m_i(\neg\{\mathrm{H_i}\}) = r_i(1 - \mathrm{L_i}) \\ m_i(\Omega) = 1 - r_i \end{cases} \tag{14}$$

where L_i is the likelihood (based on one of the four similarity measures) related to the hypothesis $\mathrm{H_i}$ that the candidate $n \in \{1, 2 \dots N\}$ corresponds to the reference object and N is the number of objects in the comparison dataset.

Although each of these studies has its own specificity, the overall steps of applying a DS-based process in object matching are quite similar. We summarize them in Figure 4. First, global/local frames of discernment are defined to identify the global/local fusion spaces according to the context of the domain application. After setting the distances between the reference object and the candidates, the goal of the transformation step is to transform these distances into similarity measures or likelihood values that are used to initialize the mass functions.

The local combination step allows to combine the mass functions locally for each couple formed by a reference object and each one of its potential candidates, based on the information given by all the considered similarity measures. The resulting mass values, combined within the global frame of discernment that contains all possible matching assumptions, make it possible to combine the information of all potentially corresponding objects. We

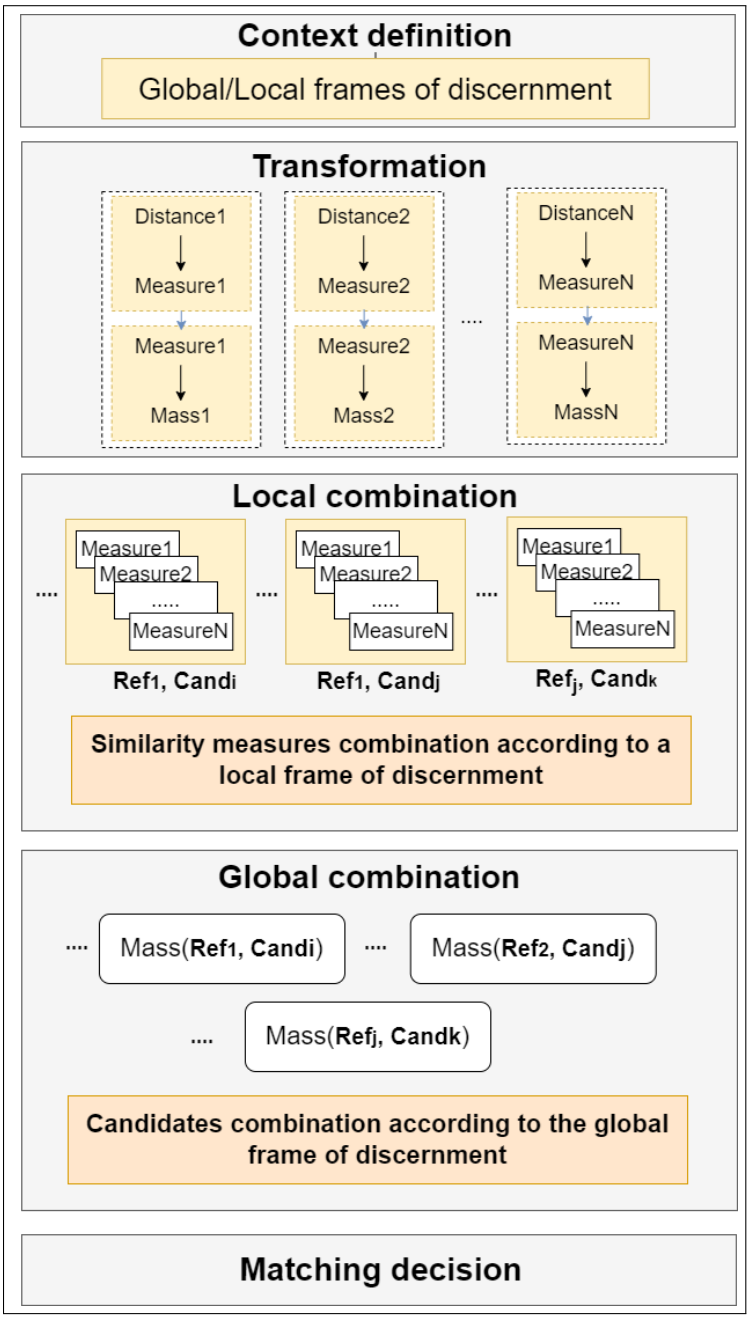


Figure 4: Main steps for applying a DS-based process in object matching

called this step global combination. Finally, a matching decision is made based on various decision tools such as the plausibility or the pignistic probability.

4 Materials and methods

4.1 General approach

Our process of matching lines representing wastewater networks is illustrated in Figure 5. As a first step, we propose to use strokes as a matching unit to overcome the constraint of missing nodes (section. 2.2). Later on, we use pipes as a matching unit for the remaining lines that do not belong to any matched stroke. The stroke detection algorithm is presented in section. 4.3.

In the second step, the purpose is to select the potential candidates for each reference object by considering two units: strokes and pipes. First, the stroke candidates are selected for each line in the reference dataset. The pipe candidates' selection only applies to lines of strokes that could not find a match during step 3. The stroke/line selection is based on a set of distances that we introduce in the section. 4.2.

In the third step, we propose an enhanced DS-based process to combine the similarity measures for stroke/line matching.

As for the partial matching, and since wastewater networks are often expanded and repaired, the remaining unmatched lines include not only the pipes that don't have a corresponding candidate, but also misrepresented and newly laid pipes. We propose to handle situations like the one illustrated in Figure 2, where a line can partially match with another line, by adding a set of uniform fictitious nodes to achieve a partial matching. However, since we cannot identify the origin of the differences in lengths, which could be due to missing nodes, replaced pipes, or errors of representation, we also increase the uncertainty of the matching when these fictitious nodes are added.

4.2 Applied Distances

In order to assess the similarity between objects of different datasets, we adopt four geographical and topological distances that are relevant to the context of the wastewater application. From each distance, a similarity measure is derived. We use:

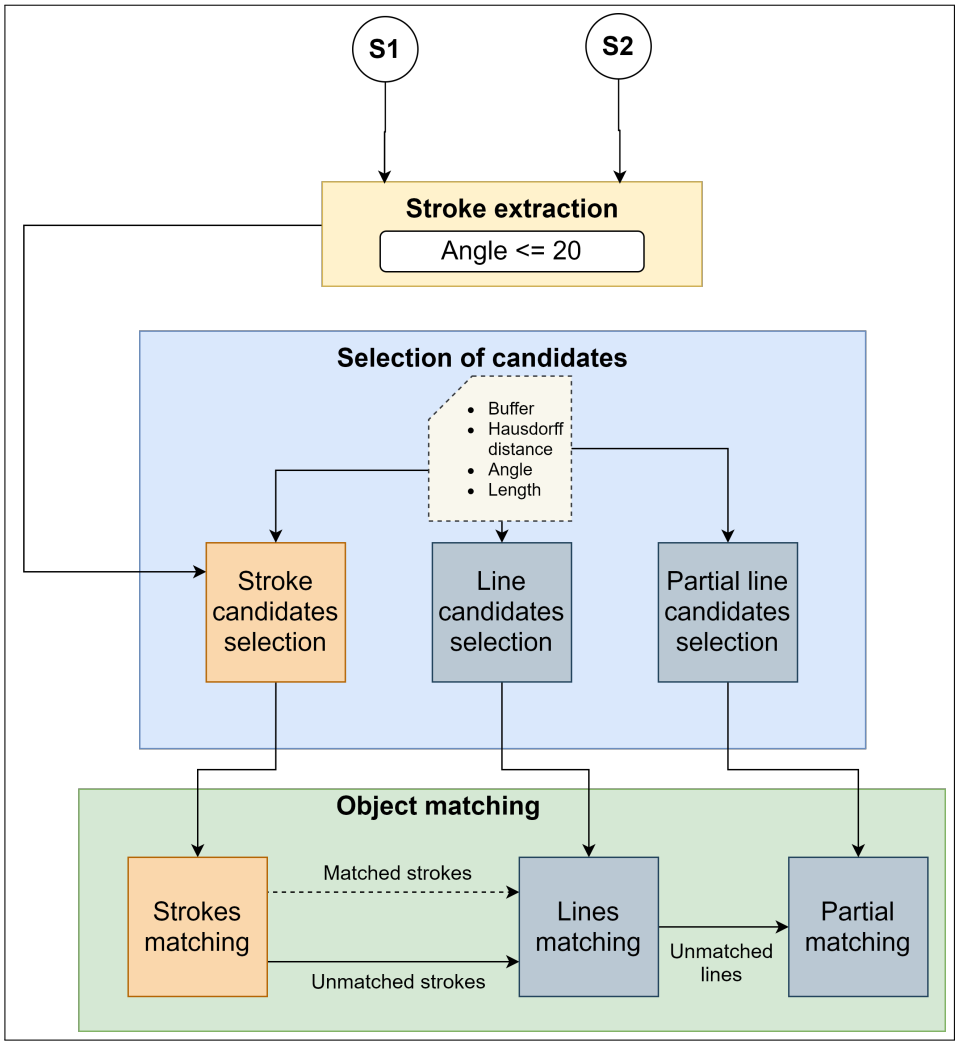


Figure 5: The proposed process for matching wastewater network pipes

- The Hausdorff distance [46]. It is a straightforward distance between two subsets of a metric space (the lines in our case), which is widely applied to linear object matching. We did not use the Fréchet distance since it is more suited for curves than lines. In addition, an outlying part of a line may be due to an extension or a replacement of the pipe. To avoid ignoring the outlying parts, we did not use the extended Hausdorff distance. Like for the Euclidean distance, the assumption is that two lines are a potential match if the Hausdorff distance between them is small. For two lines L_1 and L_2, the Hausdorff distance is defined as:

$$d_{\text{Hausdorff}}(L_1, L_2) = \max\{\min_{a \in L_1, b \in L_2} d_{\text{Euclidean}}(a, b), \min_{a \in L_1, b \in L_2} d_{\text{Euclidean}}(b, a)\} \tag{15}$$

- The Length-based distance. It is computed as the ratio of the differences in lengths, since the corresponding pipes must have similar lengths:

$$d_{\text{length}}(L_1, L_2) = \frac{|\text{length}(L_1) - \text{length}(L_2)|}{\max(\text{length}(L_1), \text{length}(L_2))} \tag{16}$$

- The Orientation-based distance. The corresponding pipes should also have similar orientations, and a strong deviation between lines may indicate that they are unlikely to correspond. The orientation of a line $L_{a,b}$, where a and b are the end nodes of L, is defined as the angle between $L_{a,b}$ and the x axis.

- The Node Degree-based distance. Node degree is a topological distance, defined for a node N as the number of lines connected to N. We use this distance to check whether the nodes to match have the same neighbourhood structure. The degree distance in terms of the node degree between two nodes n_1 and n_2 is computed as follows:

$$d_{\text{degree}}(n_1, n_2) = 1 - \frac{|\text{degree}(n_1) - \text{degree}(n_2)|}{\max(\text{degree}(n_1), \text{degree}(n_2))} \tag{17}$$

  For lines, the degree-based distance between the closest end nodes of the two lines $L_{a,b}$ and $L_{c,d}$, is first computed separately. The mean of the degree-based distance of the two couples is retained as the node

degree-based distance between the two lines:

$$d_{\text{degree}}(L_{a,b}, L_{c,d}) = \frac{d_{\text{degree}}(a,c) + d_{\text{degree}}(b,d)}{2}, \text{where} \quad d_{\text{Euclidean}}(a,c) <= d_{\text{Euclidean}}(a,d) \tag{18}$$

The Hausdorff, the length, and the orientation are transformed into similarity measures following the process described in the next section (4.4.1). The node degree is used as it is defined below as a similarity measure.

4.3 Stroke detection algorithm

In the stroke detection algorithm (see algorithm 1), we consider two connected pipes as part of the same stroke when their angles differ by less than a certain threshold, which represents the level of imperfections that we tolerate. According to a domain expert, the value of 20 degrees seems to be a good compromise. Indeed, a lower value will result in very few identified strokes (a very small set of pipes is considered as part of a stroke), whereas too high a value will lead to complex shapes unlikely to be matched. We repeat this operation until all the lines are processed. The line neighbours are the lines that have a common node.

4.4 The enhanced DS-based process

Our object matching is performed according to a DS-based process, allowing us to combine the distances in order to decide whether two or more objects match. Considering two independent sources including N and M objects respectively, we define the frames of discernment for the local measures' combination as $\Omega_l = \{\mathrm{H_{i,j}}, \neg \mathrm{H_{i,j}}\}$, where $\mathrm{H_{i,j}}$ is the hypothesis that the object $i \in \{1 \dots N\}$ from source 1 matches with the object $j \in \{1 \dots M\}$ from source 2. $\Omega = \{\mathrm{H_{1,1}} \dots \mathrm{H_{N,M}}\}$ is the global frame of discernment of the candidates' combination step. We propose to enhance the DS based process of Figure 4 according to three main contributions: introducing i) candidates' ranking and ii) a mixed model in the transformation step as well as, iii) adding an intermediate combination step after the mass initialisation that we call bidirectional measure combination (Figure 6)

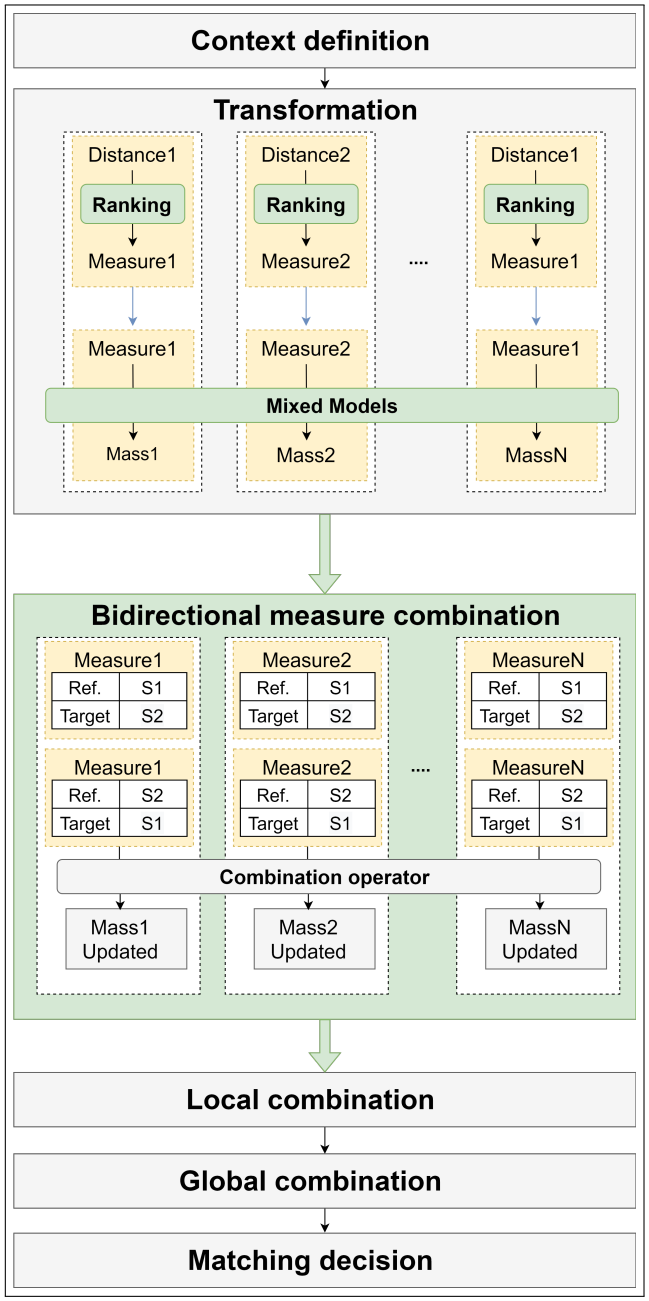


Figure 6: Proposed steps of the enhanced DS process for object matching

Algorithm 1 Stroke detection

```
Initialise strokeDict = {}
i ← 0
for each line ∈ Source do
    if line ∉ strokeDict then
        strokeDict[line] ← i
        i ← i + 1
    end if
    for each neighbor ∈ line.neighbours do
        if neighbor ∉ strokeDict then
            θ ← angle(line, neighbor)
            if |θ| ≤ 20 then
                strokeDict[neighbor] ← strokeDict[line]
            end if
        end if
    end for
end for
```

4.4.1 Candidates' ranking

The first phase of the transformation step consists in transforming distances into similarity measures. Buffers/filters are used to reduce the number of potential candidates (section. 2.3.2), but using them may not be sufficient when the distances are almost equal, since the similarity measures will be as well. In this case, candidates' ranking is important, where the closest candidate should be emphasised over the next closest one. This piece of information is relevant and should be exploited in the matching process to highlight the closest candidates and to avoid that similar distances become contradictory masses and generate an important conflict after the combination step. To our knowledge, candidates' ranking has never been considered in DS-based object matching. Nevertheless, ranking of candidates has been proposed in [34] for node matching, where a probability value is derived from the Euclidean distance by defining a parameter that decreases the similarity measure when the distance to the reference object increases. From the same perspective, we propose to rank the candidates in the first transformation

step using the following equation:

$$\text{Similarity}_{\text{metric}}(\text{H}_{\text{i,j}}) = \frac{d_{\text{metric}}(\text{H}_{\text{i,j}})^{-\beta}}{\sum_{k=1}^{\text{N}_\text{c}} d_{\text{metric}}(\text{H}_{\text{i,k}})^{-\beta}} \tag{19}$$

where N_c is the number of candidates, d_{metric} is a distance such as Hausdorff, $\text{H}_{\text{i,j}}$ is the hypothesis that the object i from source 1 matches with object j from source 2, β is a penalty factor and candidate $k \in 1 \dots |\text{candidates}|$. The more β increases, the more the gap between the closest and the remaining candidates increases. The choice of β is subjective and can vary between measures. We set $\beta = 2$ for the Hausdorff distance, and $\beta = 1$ for the other distances, as we wish to emphasize more the order of the candidates in terms of Hausdorff distance.

4.4.2 Mixed models

The second phase of the transformation step aims to transform the similarity measures into masses. As described in section 3.2, [22] proposed two different models to directly assign masses from measures to each focal element (Figure 7). Given a computed likelihood L_i, between a reference object and a potential candidate i, and a discounting factor r_i related to the reliability of the source or the measure (set to 1 when reliable), the first model initializes the mass value of the hypothesis H_i that the candidate i corresponds to the reference object to 0, regardless of the value of L_i. In addition, when the two objects are close (L_i is close to 1), this model also assigns a mass close to 0 to $\neg\text{H}_\text{i}$. We consider it "the cautious model" since it assumes that the measure at disposal represents only a single part of the reality, and it prefers not to support one hypothesis over another. On the other side, the second model is more incautious and assigns the likelihood value to the mass of H_i. This model is suitable for the distances that have important and decisive information for the matching process.

The models associated with the distances used in this work are summarised in Table 2. We separate them into 2 categories, as the cautious model is suitable for the first category, including Hausdorff, length, and orientation distances, and the incautious model is suitable for the node degree. Indeed, even if corresponding pipes from two sources are generally close and have similar lengths and orientations, they do not correspond when data sources have significant and non-uniform discrepancies and distortions. When the

$m_i = \begin{cases} m_i(\{\mathrm{H_i}\}) = 0 \\ m_i(\overline{\{\mathrm{H_i}\}}) = r_i(1 - \mathrm{L_i}) \\ m_i(\Omega) = 1 - r_i(1 - \mathrm{L_i}) \end{cases}$	$m_i = \begin{cases} m_i(\{\mathrm{H_i}\}) = r_i(\mathrm{L_i}) \\ m_i(\overline{\{\mathrm{H_i}\}}) = r_i(1 - \mathrm{L_i}) \\ m_i(\Omega) = 1 - r_i \end{cases}$
The cautious model	**The incautious model**

Figure 7: Models to initialise the mass function proposed in [22]

value of a measure is null (the objects are far), we are sure that the candidates do not correspond. However, when the value is equal to 1 for one of these metrics, we cannot conclude in a peremptory manner that the couples in question are a match, as the metric provides only an indication among others. That is to say that these 3 distances, considered together, do not have sufficiently important and decisive information for the matching.

The incautious model is appropriate for the node degree-based distance since the line connectivity, when similar, provides important evidence that the candidates match. It then has important and decisive information for the matching. In addition, the values of the derived similarity measure from the node degrees that are close to 1 indicate potential matches, while the low values (which may be due to data incompleteness) cannot confirm the non-matching. This is particularly true when strokes are considered as corresponding units since having different candidates with the same node degrees is not frequent.

Consequently, unlike the previous studies [24], [25], [26], where only one of the models is applied, we propose to use both models (mixed models) depending on the measure to be transformed into mass.

4.4.3 Bidirectional measure combination

Initialization of the mass functions is carried out based on the computed distances from a reference set to a comparison set. We propose to also compute the distances from the comparison set to the reference set and use this additional information to update the initialized masses. The values of a two-way distance for a measure may not be equal when the measure is context-aware. In our case, the ranking of candidates applies a penalty to the similarity measure, and thus, the measure values in the two directions would not be

Table 2: The models associated with the applied similarity measures

Distance	Hausdorff	Length	Orientation	Node Degree
Transformation model	Cautious Model			Incautious model

necessarily equal. In addition, with imperfections (in particular, incompleteness), the difference between the 2 way calculated measure between 2 objects is of great impact on increasing the mass value of the relevant hypothesis.

5 Experiments and results

5.1 Methodology

Like any matching proposition, our goal is to achieve the most accurate and complete set of corresponding objects. In addition, due to data imperfections, we would like to show that our approach offers the tools to model and take into account these imperfections properly. We also aim to provide a clear evaluation of the uncertainty related to the matched objects.

The objective of the experiments is to show the impact of our contributions:

1. Reducing the conflict by selecting a suitable model when transforming similarity measures to mass functions.
2. Enhancing the matching accuracy by:
    - Using strokes as matching units.
    - Considering the candidates' ranking when transforming the distances into measures.
    - Applying bidirectional combination.
    - Performing partial matching.

We first conducted tests on synthetic data, then on real-world datasets. Indeed, synthetic data allow to focus on very specific scenarios to show the impact of each of our contributions individually, before assessing them against real-world datasets.

Five configurations (summarized in table 3) were defined for 3 experiments:

Table 3: Summary of the configurations.

	Config.1	Config.2	Config.3	Config.4	Config.5 (our approach)
Ranking			X	X	X
Cautious model (model 1)		X		X	X
Incautious model (model 2)	X		X		X
Bidirectional					X

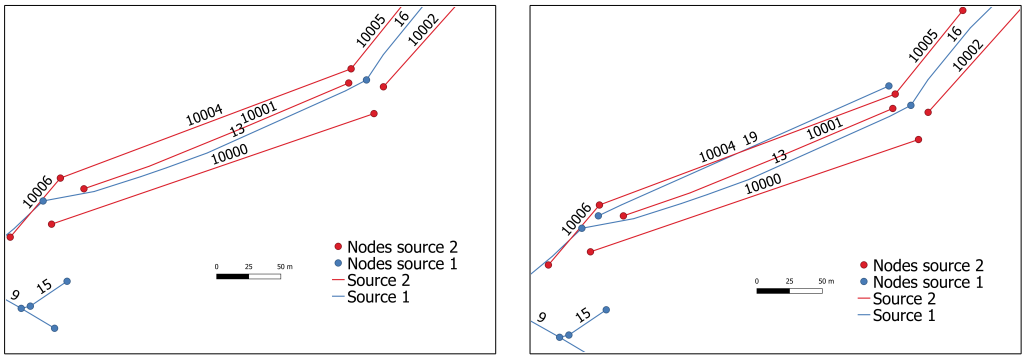
- Configurations 1 and 2 use, respectively, the cautious and the incautious models in transforming the similarity measures to mass functions.
- Configurations 3 and 4 additionally support the candidates' ranking.
- Configuration 5 supports the 3 contributions: candidates' ranking, the mixed model, and the bidirectional measure combination. It corresponds to our proposed approach.

The 3 experiments summarized in Table 4 are as follows:

- The 1st experiment is applied to the synthetic data for the Hausdorff distance only, in one direction of matching, without considering the bidirectional measure combination and the local measures' combination steps. The aim is to evaluate the impact of the cautious/incautious models on the conflict and the result of the combination step.
- The 2nd experiment is applied to the synthetic data for the Hausdorff, length, orientation, and the node degree-based measures, along with the bidirectional measure combination step. In addition, we consider that the data sources are not always reliable and apply a discount based on the source's reliability. The aim is to assess the impact of the bidirectional combination on the correctness of the matching decision. In this experiment, we compare the results of configuration 5 to configuration 4.
- The 3rd experiment is applied to the real-world datasets. The results are described and discussed based on our approach only (configuration 5).

Table 4: Summary of the three experiments.

	Dataset	Distances	Reliability	Measures Combination	Bidirectional measure	Configurations
Experiment 1	Synthetic	Hausdorff	1	No	No	Conf1, Conf2, Conf3 and Conf4
Experiment 2	Synthetic	Hausdorff, Length, Orientation and Node degree	0.8	Yes	Yes	Conf4 and Conf5
Experiment 3	Real-world	Hausdorff, Length, Orientation and Node degree	0.8	Yes	Yes	Conf5



(a) Use case 1.

(b) Use case 2.

Figure 8: Synthetic use cases.

5.2 Datasets

5.2.1 Synthetic data

Figure 8 shows lines delimited by their associated nodes from two sources (source 1 in blue and source 2 in red). We consider two use cases in Figures 8(a) and 8(b). Table 5 shows the corresponding Hausdorff distance in meters between the lines of the two sources.

The two use cases differ by the presence of the object identified by '19'

Table 5: Hausdorff distance between lines of sources 1 and 2

Lines	10000	10001	10004
13	26.85	33.16	22.33
19	46.92	19.17	8.34

in source 1 in the second use case. After creating a buffer and applying a Hausdorff distance filter, we consider the reference objects identified as '13' and '19' as having three different candidates from source 2, identified respectively as '10000', '10001' and '10004'. Based on Figure 8(a), taking into consideration the Hausdorff distance (table 5) and the node degrees of the end nodes of the lines, the corresponding objects should, theoretically, be the couples ('13','10004') and ('19','10001').

5.2.2 Real-world datasets

We consider two real-world datasets from Prades-le-Lez (Figure 9), a small city of 5,908 inhabitants [47] located in the south of France. We obtained two different datasets of the wastewater network, created by distinct organisations at different times (2014 and 2017), both provided by the managers of the network. They contain 804 and 883 pipes, representing respectively 23 km and 25.5 km of pipes. First, discrepancies in representation can be noticed in Figure 10. Second, although the 2017 network is more recent and contains more pipes, we can see that both datasets have missing pipes. Third, the nodes that represent the position of the junction of two or more pipes are indicated in detail in the 2017 dataset (1088 nodes) compared to the 2014 dataset (834 nodes).

5.3 Reducing the conflict

5.3.1 Experiment 1, Use case 1

In use case 1, the reference object is identified as '13' in source 1, and the candidates are '10000', '10001', and '10004' in source 2. The mass values for the 4 configurations are based on the Hausdorff distance between lines of sources 1 and 2 (reported in table 5). The Hausdorff measure is considered to be totally reliable (i.e., $r_i = 1$). As illustrated in Figure 11, the mass values are initialized according to the incautious model for configurations 1 and 2 and according to the cautious model for configurations 3 and 4.

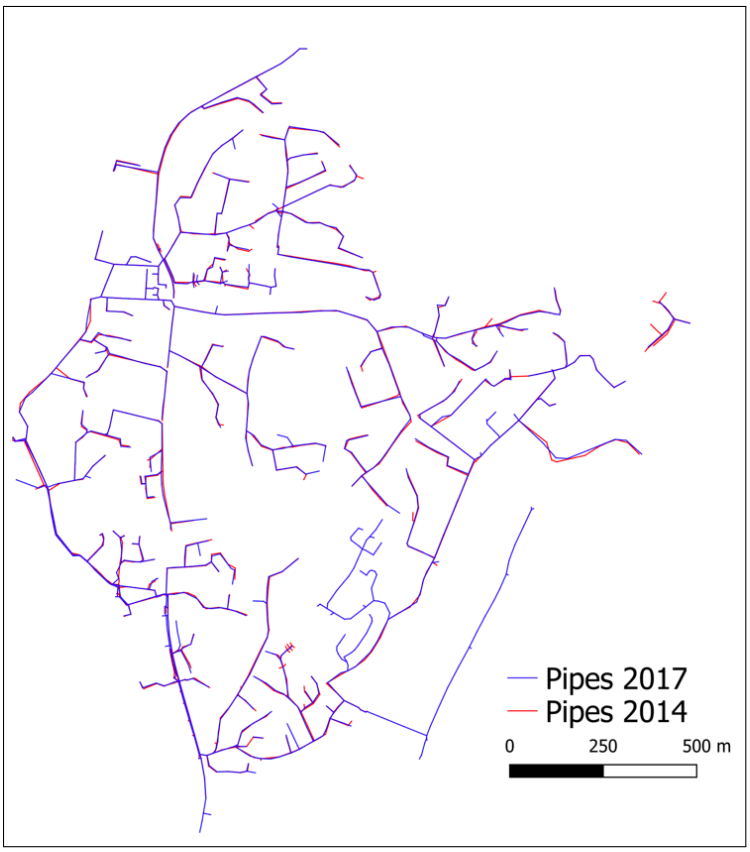
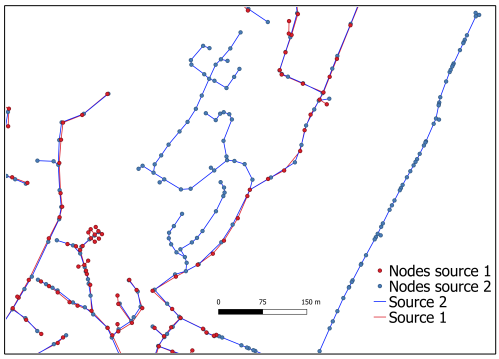
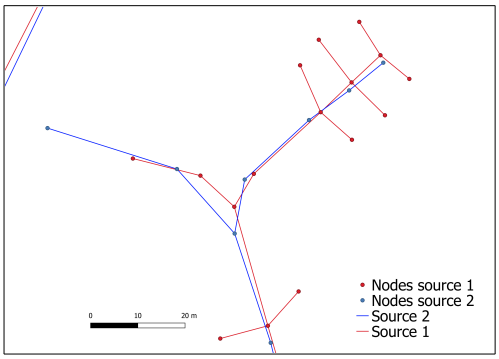


Figure 9: Maps of Prades le Lez's wastewater network in 2014 and 2017.

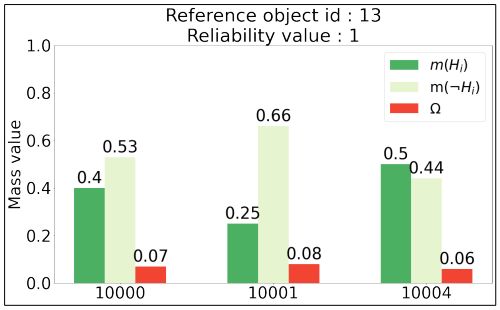


(a) Example of missing pipes in source 1.

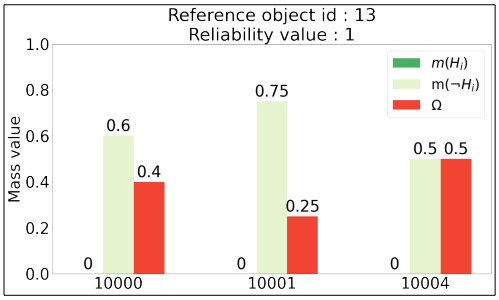


(b) Example of missing pipes in source 2.

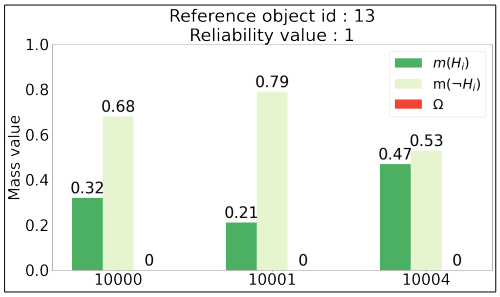
Figure 10: Examples of data imperfections in Prades-Le-Lez's datsets.



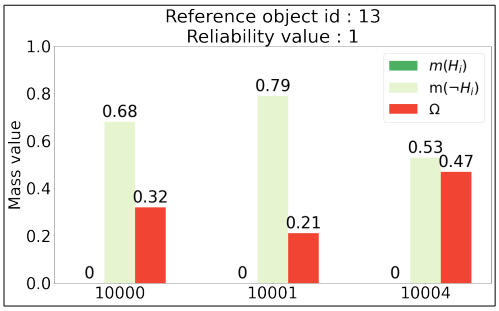
(a) Configuration 1.



(b) Configuration 2.



(c) Configuration 3.



(d) Configuration 4.

Figure 11: Initial mass values of the four configurations for the reference object '13'.

To evaluate the effect of the models on the conflict K generated by the combination of the masses, we used the conjunctive rule (equation 6), i.e., without normalizing by the conflict value. The mass values of the subsets generated after the combination steps are not displayed for the sake of clarity. Figure 12 shows the mass values after the candidates' combination step.

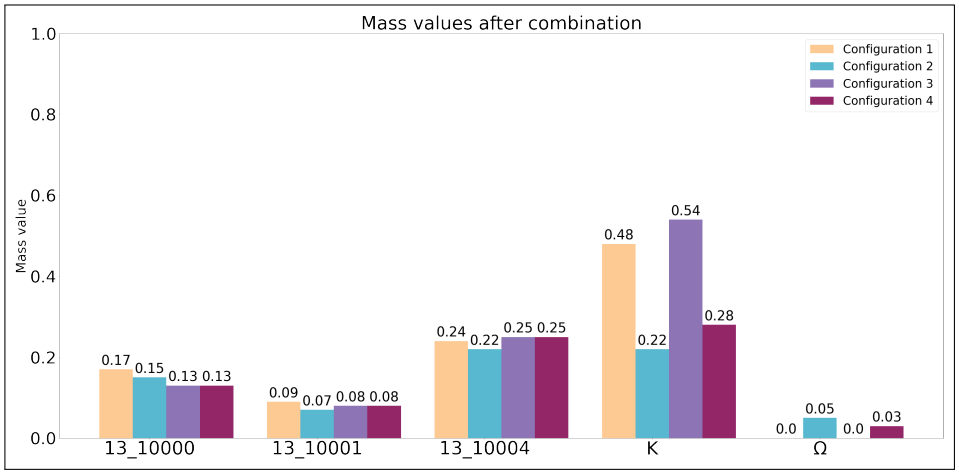


Figure 12: Experiment 1, Use case 1: The mass values obtained after the combination process.

We notice that for configurations 1 and 3, where the incautious model is applied, the values of the generated conflict are almost twice those of configurations 2 and 4. The four configurations predict the correct couple ('13', '10004') to be matched based on the plausibility measure (Figure 13), which is often used as a decision criterion. However, the plausibility values are substantially lower for configurations 1 and 3 compared to configurations 2 and 4. In use case 1, these differences are important while having a frame of discernment Ω = ('13','10000'), ('13','10001'), ('13','10004') with 3 possible couples in total.

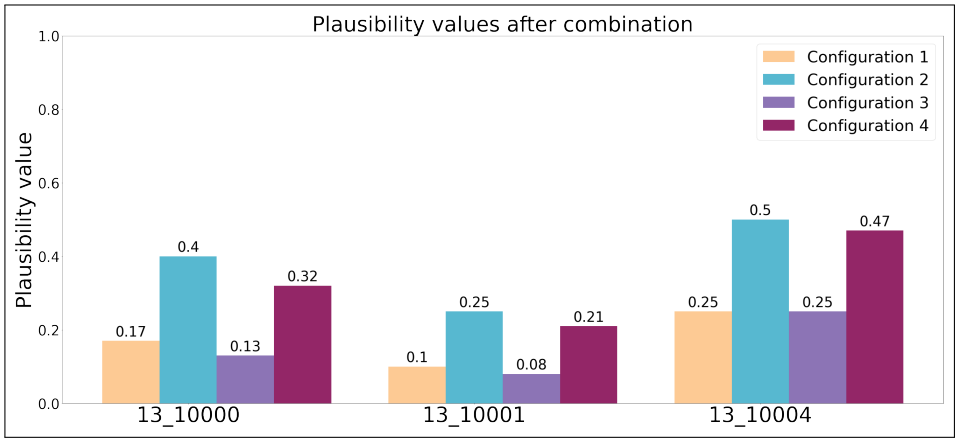
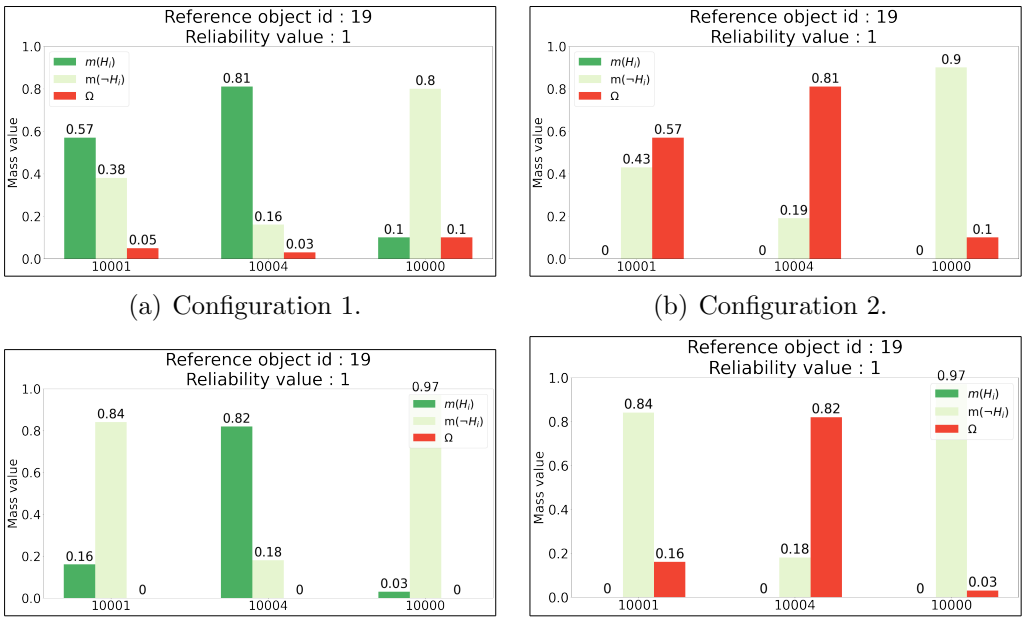


Figure 13: Experiment 1, Use case 1: The plausibility of the corresponding couples after the combination process.

5.3.2 Experiment 1, Use case 2

Use case 2 (Figure 8(b)), is based on use case 1 with an additional object identified as '19' in source 1. The frame of discernment is then defined by $\Omega = \{('13','10000'), ('13','10001'), ('13','10004'), ('19','10000'),'19','10001'), ('19','10004')\}$. Figure 14 shows the mass values' initialization for the reference object '19' in the four configurations.



(a) Configuration 1.

(b) Configuration 2.

(c) Configuration 3.

(d) Configuration 4.

Figure 14: Initial mass values of the four configurations for the reference object '19'.

The masses obtained after the combination step are illustrated in Figure 15.

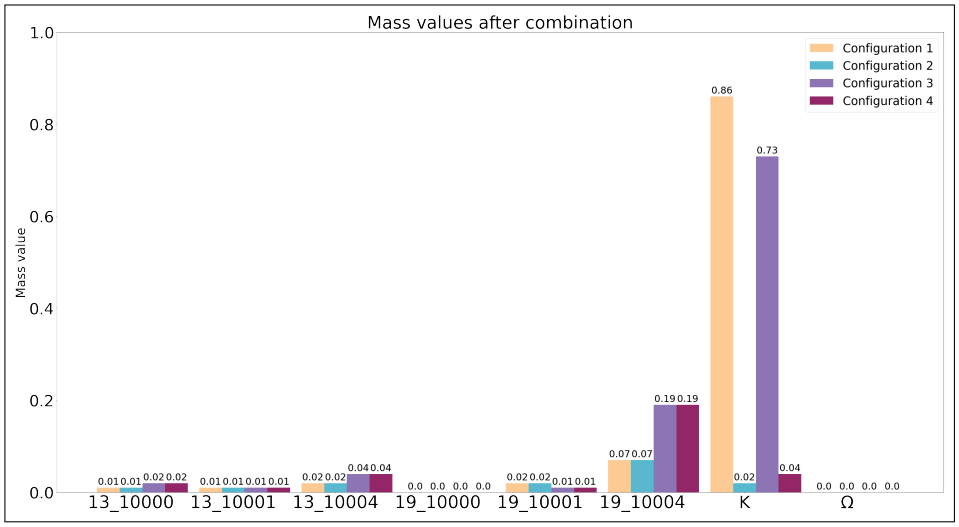


Figure 15: Experiment 1, Use case 2: The mass values obtained after the combination process.

In use case 2, the generated conflict is considerably higher than in use case 1 for configurations 1 and 3 compared to configurations 2 and 4. Also, when taking into consideration the candidates' ranking, the mass value of the closest candidate for configurations 3 and 4 is almost twice that of configurations 1 and 2 (Figure 15). In addition, the plausibility values (Figure 16) are all almost equal to zero for configurations 1 and 3, in clear contrast to configurations 2 and 4.

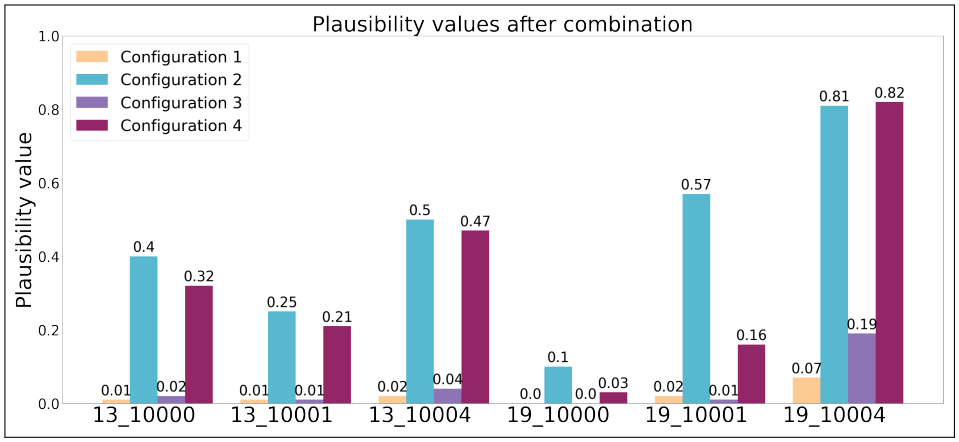


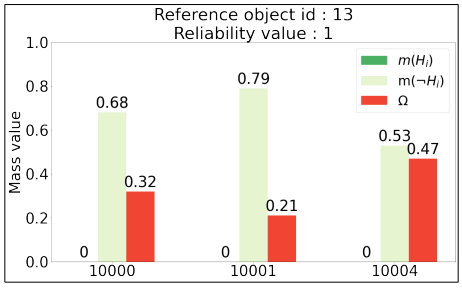
Figure 16: Experiment 1, Use case 2: The plausibility of the corresponding couples after the combination process.

These results show that the model used to transform distances to mass functions plays a key role in the final result of the combination. An important conflict is generated using a frame of discernment of small size. In real use cases where the frame of discernment may include more than 10 couples, the use of Model 2 generally results in total conflict (value equal to 1). Even if normalizing the combination will increase the mass and the plausibility values, decisions should not be made when such conflict is encountered and plausibility values are so close to 0, as shown in Figure 16. Using the Hausdorff distance only, the couples designated to match based on the plausibility values (Figure 16) are ('19','10004') and ('13','10000'). This incorrect matching is to be expected since, at this stage, only one measure is applied. In the following, we consider all the measures and compare configuration 4 to configuration 5.

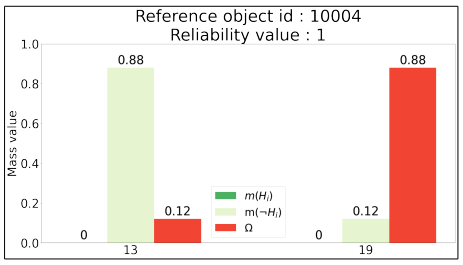
5.4 Matching accuracy

5.4.1 Experiment 2, Use case 2

Table 5 shows that the closest object to the pipe identified as '13' is the one identified as '10004'. However, when the direction of the matching is reversed by considering source 2 as reference, object '13' is not the closest object to pipe '10004'. This important piece of information can be accounted for in



(a) Mass values of the reference object identified as '13' in source 1.



(b) Mass values of the reference object identified as '10004' in source 2.

Figure 17: Initial mass values in both directions of the matching for the objects '13' and '10004'.

the matching process in the bidirectional measure combination step.

Figure 17 shows the mass values' initialization for the couple ('13','10004') in both directions of the matching. Figure 18, shows the mass values after combination of both directions (Figures 17(a) and 17(b)), where the mass $m(\neg 13\_10004)$ is no longer equal to 0.53, as it was considered above (Figure 17(a)), but equal to 0.94, thus impacting the output values of the matching.

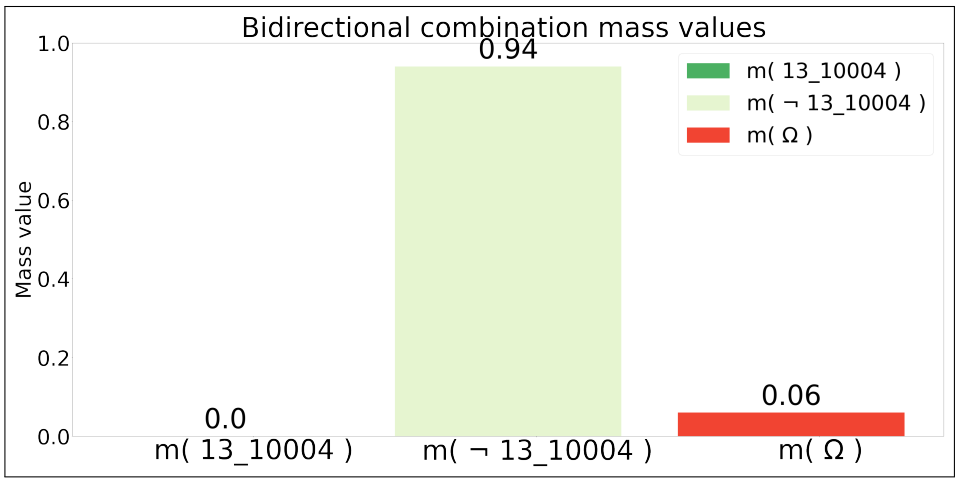


Figure 18: Experiment 2, Use case 2: Bidirectional combination step of the Hausdorff-based mass values for the couple ('13','10004').

Figure 19, shows the plausibility values when using our matching process based on the normalized conjunctive rule of combination and a reliability value of 0.8 for all the measures. By using the mixed models and the bidirectional combination, we obtain the correct corresponding couples $('13', 10004')$ and $('19', 10001')$. Hence, the use of bidirectional matching can impact the final results greatly. In this case, it leads to the correct corresponding couples while choosing the suitable model yields a reasonable conflict.

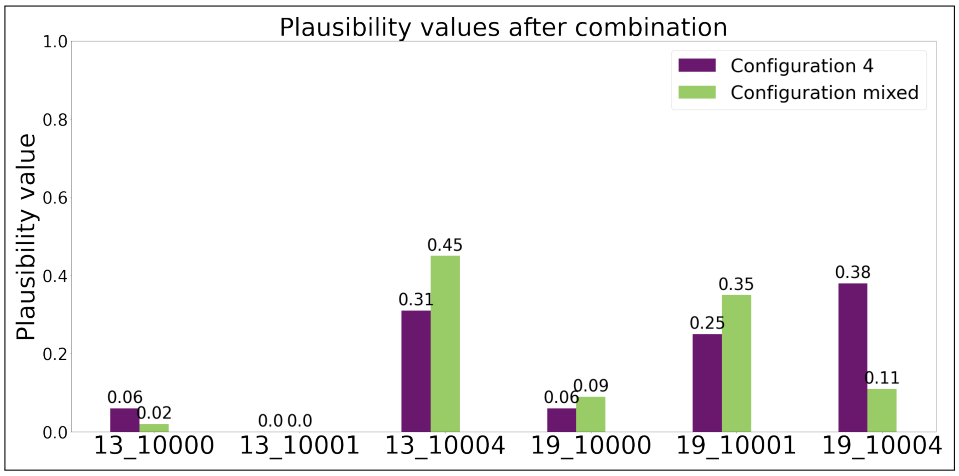


Figure 19: Experiment 2, Use case 2: The plausibility values after applying our matching process for Configurations 4 and 5 (mixed).

5.4.2 Experiment 3 (Real-world datasets)

We applied our proposed approach (configuration 5) to automatically find corresponding objects between the two datasets of Prades-Le-Lez. We refer to the 2014 and 2017 datasets, respectively as source 1 and source 2. The stroke construction step resulted in 398 and 421 strokes, respectively. After extracting the strokes and identifying potential candidates for each stroke, our enhanced DS-based combination process (Figure 6) was carried out. Due to the imperfections of the sources, we combined the four similarity measures with a reliability of 0.8 for each, thus leaving room for ignorance.

At this stage, 277 strokes were matched (Table 6). They represent 576 pipes (71.6%) of source 1 with a total length of 17.296 km and 580 pipes (65.7% and 17.301 km) of source 2. Among them, 14 strokes were falsely

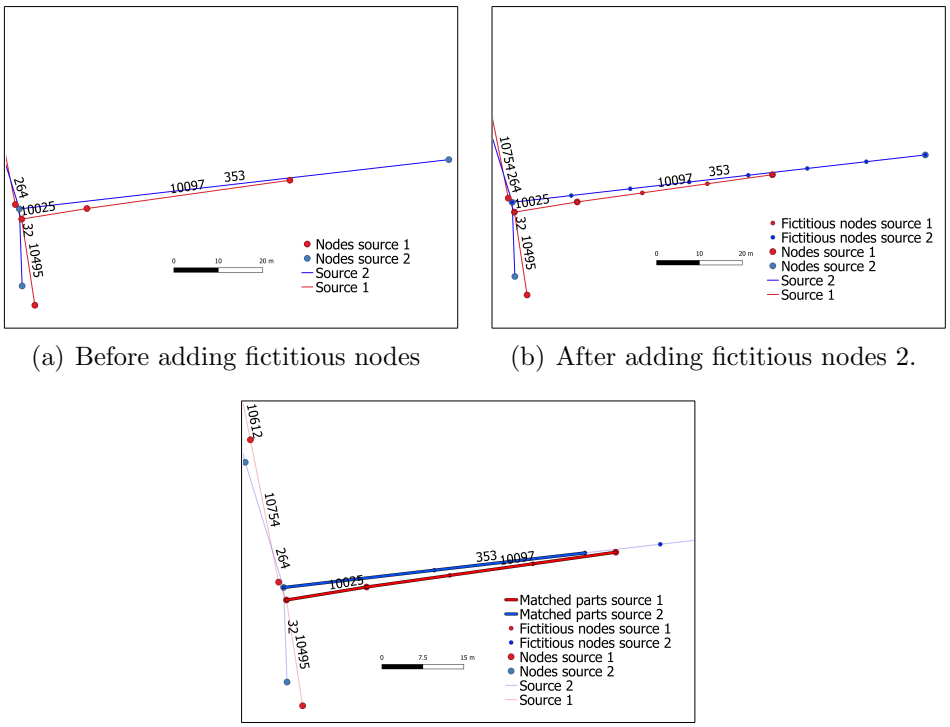
Table 6: The matching results for each step using our proposition on real-world data.

Enhanced DS-based matching	Matched strokes	Matched pipes (Source 1-2014)	Matched pipes (Source 2-2017)
Stroke matching	277	576 (17.296km)	580 (17.301km)
+ Line matching	277	647 (20km)	651 (20km)
+ Partial matching	277	695	697

matched, which yields a precision of 94.95%. No further operations were applied to the strokes that matched. However, an optional step of identifying corresponding lines within the matched strokes can be conducted, depending on the application (e.g., when the task is to complete missing attributes from one dataset using the other).

The enhanced DS-based process is then applied directly on the remaining pipes, which did not match using strokes as a matching unit. Considering missing nodes, we decreased the reliability of the node degree measure from 0.8 to 0.7. This results in matching 71 pipes from both sources, all being true positives. This relatively small number of matched pipes is due to missing nodes, which affect the lines' length. After this step, the total number of pipes that can be considered as corresponding is 647 for source 1 and 651 for source 2, representing almost 20 km of pipes for both sources (Table 6).

To address the length constraint due to missing nodes, we conduct a partial matching. In this step, we split the pipes into smaller sub-lines by adding new fictitious nodes. We applied the partial matching by adding nodes with a maximum spacing of 16 metres, since a high value would yield few matching couples, and a low value would generate a great number of subsets for the combination process. Figure 20, shows an example of partial matching by adding fictitious nodes. As we can see in Figure 20(a), pipes from source 1 identified as '10025' and '10097' can partially match with the pipe identified as '353' from source 2. This cannot be achieved in the previous steps, since nodes are missing. Figure 20(b), shows that no fictitious nodes were added to the pipe identified as '10025' since its length is smaller than 16 meters, whereas several nodes were added to pipes '353' and '10097'. Figure 20(c) shows the matched parts of each pipe. For the entire datasets of Prades-Le-Lez this step resulted in 48 pipes from source 1 being partially matched with 46 pipes from source 2 (Table 6).



(a) Before adding fictitious nodes

(b) After adding fictitious nodes 2.

(c) Partial matching results

Figure 20: Example of partial matching by adding fictitious nodes.

A manual matching of the two datasets has been conducted. It resulted in 731 pipes from source 1 that should have a corresponding pipe in source 2. Comparing our results to the manual matching, we found 685 true positives in the fully and partially matched pipes from source 1. We counted 16 false positive matching and 57 false negatives, that is a precision value equal to 97.7% and a recall of 92.3%. To keep track of the uncertainty related to the matched couples, the pignistic probability and the plausibility values were preserved for each matching.

These results show that both stroke transformation and partial matching help surpass the issue of missing nodes in real-world datasets. In addition, the use of DS-theory guarantees the traceability of matching uncertainties related to the matched couples.

5.5 Discussion

In this work, we aimed to use DS theory to enhance the end user's knowledge of the network structure, enabling more informed and effective decision-making. This enhanced knowledge empowers stakeholders to optimize operations, prioritize maintenance, and make proactive decisions to ensure the network's efficiency and sustainability. For instance, a clear assessment of the uncertainties in the network map makes it possible to pinpoint the areas or sections with the greatest inaccuracies or lack of information. This allows for prioritizing excavation and detection investments in these specific zones. In this section, we examine our results and the impact of our solution from the perspectives of both ***daily management*** and ***methodology***.

From ***methodological and technical perspectives***, the previous experiments were set to: i) manage/reduce the conflict, ii) enhance the matching accuracy. Indeed, we showed in the first experiment that using the inappropriate model to transform the similarity measures to mass values can lead to high conflict. The latter was almost equal to 1 in use case 2, which had only a frame of discernment of 6 hypotheses. Given the high number of candidates in real use cases, the conflict can easily get close to 1, thus rendering the model unsuitable for decisions about the corresponding objects. Mixing the models based on the semantics of each similarity measure turned out to be efficient in reducing the conflict drastically. Indeed, we have shown the impact of adopting a transformation model according to the importance of the information carried by the corresponding similarity measure, for the target application domain. The incautious model is suitable for the measures that have important and decisive information for the matching process, while the cautious model is more appropriate for lightweight measures that do not significantly influence the matching decision.

To enhance the matching accuracy, we first applied the stroke concept to address the challenge of missing nodes while conducting a matching process on lines. Indeed, when the stroke step is not taken into consideration for experiment 3, only 55% of the pipes from source 1 are directly matched before the partial process, compared to 80% of the pipes when the strokes are used. When lines are directly used as matching units, the number of candidates can be high, and several candidates for the same reference object may have similar geographic and topological structures, rendering the matching decision more difficult. The results showed that the use of candidates' ranking and the bidirectional combination succeeded in highlighting the closest candidates

and in avoiding that similar distances become contradictory masses that may generate an important conflict. Nevertheless, it is important to note that when conducting the partial matching, the distance between the fictitious nodes should be reasonable and take into consideration information from the domain knowledge, such as the minimum length of a pipe. Otherwise, it would lead to partial matching, which is practically useless for real-world purposes such as the excavation of underground networks.

We managed dataset imperfections by using the DS theory. The matching uncertainty may be measured by the mass function, the plausibility, and the pignistic probability. In addition, when the uncertainty about a matching couple is greater than a chosen threshold, one may use subsets with lower uncertainty values to indicate imprecision about the decision. For example, for experiment 1, use case 1, one may set the rule that no couple should have a plausibility value less than 0.6 (Figure 13), which can be considered as an acceptable threshold. In this case, no hypothesis can be retained as they all have values under 0.6, thus, the subset of cardinally 2 with the highest value and superior to 0.6 is $\{'13\_10000', '13\_10004'\}$ and should be considered as imprecise matching. Moreover, if newly acquired information designates a subset of the frame of discernment as the only possible couples, one may use the conditioning operation of the DS theory to propagate this information. The incompleteness is supported through the combination of local measures. That is, if both a reference and a candidate object have an attribute that is not available for the other candidates, a measure based on this attribute can be combined locally with the rest of the similarity measures. For example, given that the objects identified as '13' and '10000' respectively from source 1 and source 2 have the attribute 'diameter of pipe', a distance, then a mass function can be computed from both values and combined with the other similarity measures.

From ***the perspective of the stakeholders***, our solution can be viewed as a general monitoring tool for the management of wastewater networks where the main challenges are addressed, such as sources' heterogeneity, attribute imperfections, and missing spatial data. Our results significantly shape the daily management practices, driving improvements in three critical areas:

- Object matching: there is no longer a need to manually analyse and identify matching objects from different sources. Shape, attributes, and context are taken into account. Missing nodes are addressed using

strokes and partial matching techniques, thereby enhancing accuracy and accelerating the matching process.

- Imperfections quantification: uncertainty and imprecision are quantified and provided for each matching object, offering the end-user a clear understanding of the current available knowledge about the network. This enables budgets to be strategically allocated to reduce uncertainties in critical areas through targeted inspections. In contrast, this cannot be achieved by traditional or optimisation methods [9], [11], [12], [13].

- Data update: given the hierarchical structure of the network, manually updating it with a new available piece of data is highly complex and may require propagation across a significant number of pipes. Our DS-based solution provides the necessary tools to revise matching results when new information becomes available, enabling updates to the current knowledge accurately and efficiently.

6 Conclusion

In this work, we proposed a novel process for matching spatial objects, namely wastewater network elements. After describing the imperfections related to wastewater databases and their consequences, we presented our proposition that relies on three concepts: strokes, partial matching, and DS theory. Strokes have been applied to capture the overall structure of the networks, while partial matching has been proposed to resolve the problem of missing nodes. We adopted the DS theory to combine the similarity measures, allowing to support the uncertainty, unreliability, and imprecision in the input and output of the matching process. The particularity of our process lies in using different and suitable models to transform distances to mass functions depending on the type of information available. This resulted in reasonable conflict values after information combination, in comparison to other approaches presented in the literature review.

Although this matching process is intended for linear objects of wastewater networks, it may be used for a broader range of linear networks such as roads, rivers, and other underground networks. The DS theory process is generic and can be adopted for matching nodes or polygons, provided that suitable distances are defined. In future works, we aim to use the wide range

of operators offered by the DS theory, such as refinement and conditioning, to increase the level of detail after the matching operation by identifying the type of pipes. In addition, and after using only geographical and topological distances, we aim to exploit the semantics of the domain and regulatory rules related to the construction of wastewater networks, such as the minimum length of a pipe, to enhance the results and reduce the uncertainty of the matching.

7 Acknowledgements

This work was carried out within the framework of the CIFRE-France/Morocco Program. We thank Mustapha Derras for the general supervision of the research group and administrative support. We thank Montpellier Métropole Mediterrannée for the wastewater network data.

This research has also received support from the European Union's Horizon research and innovation program under the MSCA-SE (Marie Skłodowska-Curie Actions Staff Exchange) grant agreement 101086252; Call: HORIZON-MSCA-2021-SE-01; Project title: STARWARS (STormwAteR and WastewAteR networkS heterogeneous data AI-driven management). And from the French national project ANR (Agence Nationale de la Recherche) CROQUIS (Collecte, représentation, complétion, fusion et interrogation de données hétérogènes et incertaines de réseaux d'eaux urbains).

References

[1] Wout Broere. Urban underground space: Solving the problems of today's cities. Tunnelling and Underground Space Technology, 55:245–248, 2016.

[2] INERIS. Ineris. https://www.reseaux-et-canalisations.ineris.fr/gu-presentation/construire-sans-detruire/prevenir-les-risques.html, 2022. Accessed 20 september 2022.

[3] Ahmed Jalil Al-Bayati and Louis Panzer. Reducing damage to underground utilities: Lessons learned from damage data and excavators in north carolina. Journal of Construction Engineering and Management, 145(12):04019078, 2019.

[4] USAG Data and Reporting Working Group. Utility Strike Damages Report, 2019.

[5] Congressional Research Service. DOT's Federal Pipeline Safety Program: Background and Key Issues for Congress, 2022.

[6] Phil Goodwin. Utilities' street works and the cost of traffic congestion, 2005.

[7] L'institut national de recherche et de sécurité (INRS). Travaux à proximité des réseaux enterrés et investigations complémentaires sans fouille, 2014.

[8] ASTEE. Gestion patrimoniale des réseaux d'assainissement, 2015.

[9] Thomas Devogele, Christine Parent, and Stefano Spaccapietra. On spatial database integration. International Journal of Geographical Information Science, 12(4):335–352, 1998.

[10] Xiaohua Tong, Wenzhong Shi, and Susu Deng. A probability-based multi-measure feature matching method in map conflation. International Journal of Remote Sensing, 30(20):5453–5472, 2009.

[11] Steffen Volz. An iterative approach for matching multiple representations of street data. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 36(Part 2/W40):101–110, 2006.

[12] Wenbo Song, James M Keller, Timothy L Haithcoat, and Curt H Davis. Relaxation-based point feature matching for vector map conflation. Transactions in GIS, 15(1):43–60, 2011.

[13] Zhen Lei, Zhangshun Yuan, and Ting L Lei. On the theoretical link between optimized geospatial conflation models for linear features. ISPRS International Journal of Geo-Information, 13(9):310, 2024.

[14] David L Hall and James Llinas. An introduction to multisensor data fusion. Proceedings of the IEEE, 85(1):6–23, 1997.

[15] Duncan Smith and Sameer Singh. Approaches to multisensor data fusion in target tracking: A survey. IEEE transactions on knowledge and data engineering, 18(12):1696–1710, 2006.

[16] Haoran Duan, Jiuling Li, and Zhiguo Yuan. Making waves: Knowledge and data fusion in urban water modelling. Water Research X, 24:100234, 2024.

[17] Qinli Zhang, Pengfei Zhang, and Tianrui Li. Information fusion for large-scale multi-source data based on the dempster-shafer evidence theory. Information Fusion, 115:102754, 2025.

[18] Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. IEEE Transactions on Multimedia, 25:3420–3431, 2022.

[19] Moxian Song, Chenxi Sun, Derun Cai, Shenda Hong, and Hongyan Li. Classifying vaguely labeled data based on evidential fusion. Information Sciences, 583:159–173, 2022.

[20] Liguo Fei, Tao Li, and Weiping Ding. Dempster–shafer theory-based information fusion for natural disaster emergency management: A systematic literature review. Information Fusion, page 102585, 2024.

[21] Zhongyi Michael Zhang, Sofia Catalucci, Adam Thompson, Richard Leach, and Samanta Piano. Applications of data fusion in optical coordinate metrology: a review. The International Journal of Advanced Manufacturing Technology, 124(5):1341–1356, 2023.

[22] Alain Appriou. Uncertain data aggregation in classification and tracking processes. Springer, 1998.

[23] Alain Appriou. Uncertainty theories and multisensor data fusion. John Wiley & Sons, 2014.

[24] Yue Deng, An Luo, Jiping Liu, and Yong Wang. Point of interest matching between different geospatial datasets. ISPRS International Journal of Geo-Information, 8(10):435, 2019.

[25] Ana-Maria Olteanu Raimond, Sébastien Mustière, and Anne Ruas. Knowledge formalization for vector data matching using belief theory. J. Spatial Inf. Sci., 10:21–46, 2015.

[26] Ghalia Nassreddine, Fahed Abdallah, and Thierry Denoeux. Map matching algorithm using interval analysis and dempster-shafer theory. In 2009 IEEE Intelligent Vehicles Symposium, pages 494–499. IEEE, 2009.

[27] Didier Dubois and Henri Prade. Formal representations of uncertainty. In Bouyssou D, Dubois D, Pirlot M, Prade H, eds. Decision-making process-concepts and methods, pages Chapter 3. 85–156. ISTE & Wileyr, 2009.

[28] Shuai Wang, Qingsheng Guo, Xinglin Xu, and Yuwu Xie. A study on a matching algorithm for urban underground pipelines. ISPRS International Journal of Geo-Information, 8(8):352, 2019.

[29] Robert C Thomson and Rupert Brooks. Exploiting perceptual grouping for map analysis, understanding and generalization: The case of road and river networks. In International Workshop on Graphics Recognition, pages 148–157. Springer, 2001.

[30] Linna Li and Michael F Goodchild. An optimisation model for linear feature matching in geographical data conflation. International Journal of Image and Data Fusion, 2(4):309–328, 2011.

[31] Benoît Costes. Matching Old Hydrographic Vector Data from Cassini's Maps. e-Perimetron, 9(2):51–65, 2014.

[32] Meng Zhang and Liqiu Meng. An iterative road-matching approach for the integration of postal data. Computers, Environment and Urban Systems, 31(5):597–615, 2007.

[33] Bin Jiang, Sijian Zhao, and Junjun Yin. Self-organized natural roads for predicting traffic flow: a sensitivity study. Journal of statistical mechanics: Theory and experiment, 2008(07):P07008, 2008.

[34] Catriel Beeri, Yaron Kanza, Eliyahu Safra, and Yehoshua Sagiv. Object fusion in geographic information systems. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, pages 816–827, 2004.

[35] Yanxia Wang, Deng Chen, Zhiyuan Zhao, Fu Ren, and Qingyun Du. A back-propagation neural network-based approach for multi-represented feature matching in update propagation. Transactions in GIS, 19(6):964–993, 2015.

[36] Jung Ok Kim, Kiyun Yu, Joon Heo, and Won Hee Lee. A new method for matching objects in two different geospatial datasets based on the geographic context. Computers & Geosciences, 36(9):1115–1122, 2010.

[37] Ashok Samal, Sharad Seth, and Kevin Cueto 1. A feature-based approach to conflation of geospatial sources. International Journal of Geographical Information Science, 18(5):459–489, 2004.

[38] Emerson MA Xavier, Francisco J Ariza-López, and Manuel A Urena-Camara. A survey of measures and methods for matching geospatial vector datasets. ACM Computing Surveys (CSUR), 49(2):1–34, 2016.

[39] Helmut Alt, Christian Knauer, and Carola Wenk. Comparison of distance measures for planar curves. Algorithmica, 38(1):45–58, 2004.

[40] Deng Min, Li Zhilin, and Chen Xiaoyong. Extended hausdorff distance for spatial objects in gis. International Journal of Geographical Information Science, 21(4):459–475, 2007.

[41] Alan Saalfeld. Conflation automated map compilation. International Journal of Geographical Information System, 2(3):217–228, 1988.

[42] Ali Assi and Wajdi Dhifli. Instance matching in knowledge graphs through random walks and semantics. Future Generation Computer Systems, 123:73–84, 2021.

[43] Meng Zhang, Wei Shi, and Liqiu Meng. A generic matching algorithm for line networks of different resolutions. In Workshop of ICA commission on generalization and multiple representation computering faculty of a Coruña University-Campus de Elviña, Spain, volume 9, pages 101–110. Citeseer, 2005.

[44] Philippe Smets and Robert Kennes. The transferable belief model. Artificial intelligence, 66(2):191–234, 1994.

[45] Alain Appriou. Probabilities and unknowns in multisensor data fusion(probabilites et incertitude en fusion de donnees multi-senseurs). Revue Scientifique et Technique de la Defense, 1 st Quarter, 1991,, pages 27–40, 1991.

[46] William Rucklidge. Efficient Visual Recognition Using the Hausdorff Distance. Springer-Verlag, Berlin, Heidelberg, 1996.

[47] INSEE, 2019.