Categorizing data imperfections for object matching in wastewater networks using belief theory

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

Categorizing data imperfections for object matching in wastewater networks using belief theory

Omar Et-targuy$^{1[0000-0002-8307-7307]}$, Yassine Belghaddar$^{1,2,3,4[0000-0002-1944-6294]}$, Ahlame Begdouri$^{1[0000-0002-9967-0439]}$, Nanée Chahinian$^{2[0000-0002-0037-5377]}$, Abderrahmane Seriai^{3}, and Carole Delenne$^{2,4[0000-0001-6683-4399]}$

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

Abstract. Nowadays, data on wastewater networks covering the same geographical territory are available from different sources. The fusion of multi-source spatial data provides a new and richer dataset that can serve several purposes such as quality improvement, decision making, or delivery of new services. It has given rise to several research works focused on the visualization, analysis, and fusion of spatial databases. However, the original data is often imperfect: imprecise, uncertain, vague, incomplete, etc. Therefore, it is essential to use formalisms allowing the modeling of imperfections and to propose adapted fusion mechanisms. In this work, we aim to handle data imperfections in a generic way. We first propose a categorization, according to several dimensions, of data imperfections encountered when fusing multi-source spatial data. We then propose to model these imperfections according to the formalism of the belief theory. We consider our conducted experiments that allowed us to match nodes and edges in the different cases of data imperfection, as promising.

Keywords: Wastewater networks · Imperfection · Fusion · Belief theory.

1 Introduction

Big data treatment is a main challenge in the field of information systems. Nowadays, and with the adoption of new technologies such as Internet, smartphones, connected objects and GPS, the data related to a domain are now available from several sources with huge masses. However, this quantitative explosion has given rise to new problems related to its processing and exploitation.

Wastewater network data is available from several sources (private managers' databases, high resolution images, open data, pdf documents). The fusion of multi-source wastewater network data allows to create a new and richer data set that can help in decision making inter alia. However, most of the time, this information is imperfect which doesn't hep treating it in an homogeneous manner. Data imperfection, in the field of artificial intelligence, is processed according to the three aspects: imprecision, uncertainty and incompleteness [**olteanu2008fusion**, **olteanu2015knowledge**, **lefevre2002belief**]. It is then essential to use formalisms that allow to model these imperfections and propose adapted fusion mechanisms. The formalism of the belief theory unifies all uncertainty theories, allows to represent knowledge in a relatively natural way and enables the modeling of various forms of imperfection. This is why we have essentially turned to this formalism.

This paper is organised as follows, the first section is devoted to the research context where we focus on the main mathematical methods of belief theory and the related works. In the second section, we propose our categorization of the forms of imperfection related to wastewater networks. In section three, we expose the application of the the belief theory for wastewater object matching. Section four is devoted to the experiments and results obtained for each category of imperfection. Finally, section five concludes the paper and presents future perspectives and improvements.

2 Research Context

2.1 Concepts of the belief theory

The belief theory is also known in the literature as Dempster-Shafer theory and more particularly as evidence theory [**lefevre2002belief**]. The strength of belief functions in modeling uncertain knowledge was first demonstrated by Shafer [**shafermathematical**]. Indeed, it allows to represent imperfect data in a more natural way than with probabilities. An extension of this theory is proposed through the Transferable Beliefs Model TBM [**smets1994transferable**], which is characterized by its fundamentally non-probabilistic character. This approach separates the relation between belief representation and decision making. In this section, we briefly describe the basic concepts of the transferable belief model.

Frame of discernment The frame of discernment or frame of interest noted Ω, indicates the whole set of the possible answers (H_i) to a problem (Hypotheses):

$$\Omega = \{H_1, \ldots, H_K\} \tag{1}$$

From the frame of discernment Ω, we consider the derived set 2^Ω, including the 2^k subsets $A \subseteq \Omega$:

$$2^\Omega = \{A, A \subseteq \Omega\} = \{\{H_1\}, \{H_2\}, \ldots, \{H_K\}, \{H_1 \cup H_2\}, \ldots, \Omega\} \tag{2}$$

where 2^Ω contains the different Ω hypotheses, but also all the possible disjunctions of these hypotheses. This set allows to define the set of quantities used by the theory of belief functions to evaluate the truth of a proposition.

Basic belief assignment Given a question Q to be answered and a frame of discernment containing all of the possible solutions to this question $\Omega = \{H_1, \ldots, H_K\}$. A function m, called *basic belief assignment or bba*, also called *masse*, is defined from 2^Ω to values in the interval $[0, 1]$, with as constraint:

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{3}$$

The amount $m(A)$ represents the belief that A, as element of 2^Ω, contains the answer to the question Q. A set A such that $m(A) > 0$ is called a focal element. The mass function represents an imperfect knowledge on Ω.

Combination The combination phase allows to synthesize, in the form of a single belief function, all the knowledge coming from several functions. The objective of this step is the fusion of the complementarity and redundancy of the knowledge of different sources in order to obtain a more general knowledge, in the form of a more precise and reinforced belief function. The fundamental combination method, presented in the TBM, is the conjunctive combination rule. For two functions of masses m_1 and m_2, this rule is defined by:

$$m_1 \oplus_2 (A) = m_1 \oplus m_2(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \quad \forall A \subseteq \Omega \tag{4}$$

The mass function m corresponds then to the synthesis of the knowledge of m_1 and m_2 related to A. This conjunctive combination rule allows to verify several properties like associativity, commutativity and has a neutral element which is the empty mass function.

One of the particularities of this rule is to generate a non-normalized mass function $m(\emptyset) > 0$. As this condition is not possible in a closed world, a normalization phase is necessary. The conjunctive combination normalization is called the orthogonal combination rule or the Dempster combination rule, it is also the first rule presented by the theory and is written as follows:

$$m_{1 \oplus 2}(A) = m_1 \oplus m_2(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \quad \forall A \subseteq \Omega, A \neq \emptyset \tag{5}$$

where

$$\kappa = m_{1 \oplus 2}(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B) m_2(C). \tag{6}$$

κ is the conflict mass which takes values between 0 and 1. In general, it reflects the degree of contradiction between the combined sources. For $\kappa = 1$, $S1$ and $S2$ are entirely considered as conflicting and the sources are note fused. On the contrary, if $\kappa = 0$, the sources are in perfect agreement.

Reliability of sources When a confidence knowledge on a source S that has given a mass function m is available, it is possible to consider this meta-knowledge in order to perform a weakening operation. Let $\alpha \in [0, 1]$, a weakening coefficient, the weakened mass function is obtained by:

$$\begin{cases} \alpha_{m(\text{ A})} = (1-\alpha) m(\text{ A}), & \forall A \subset \Omega \\ \alpha_{m(\Omega)} = (1-\alpha) m(\Omega) + \alpha \end{cases} \tag{7}$$

The pignistic probability The decision phase is based on the pignistic distribution [**smets2005decision**] noted BetP and obtained from the mass function m. It is also called pignistic probability for the set of probabilities on the singletons it generates. The transfer of the mass function m into a pignistic probability function BetP expressed on Ω, is characterized for any $H_k \in \Omega$ by:

$$\operatorname{BetP}(H_k) = \frac{1}{1 - m(\emptyset)} \sum_{A \in H_k} \frac{m(A)}{|A|} \tag{8}$$

After this transformation, it is impossible to find the initial mass function m. Indeed, a mass function is only linked to a single pignistic probability but a pignistic probability can be obtained from an unlimited number of mass functions.

2.2 Related works

The belief theory is widely used in various areas. In the field of multiple criteria decision making [**beynon2000dempster**], the approach consists of constructing a belief set for each criteria to be studied. The different sets of beliefs are then combined by Dempster's combination rule to allow the decision making. The most important works of belief theory in the area of data analysis are those of T.Denoeux [**denoeux1997advanced**, **denoeux2000neural**, **denoeux2001handling**, **zouhal1998evidence**]. Various problems are discussed, including regression, form recognition and classification. This theory allows in this context to treat noisy and imprecise data. In the area of data fusion, the most important benefit results from the combination step, which generally allows to reduce the uncertainty on a prediction by using the redundancy and complementarity of the information.

2.3 The belief theory in matching road networks

Object matching is the identification of corresponding objects in different data sources [**volz2006iterative**, **tong2009probability**]. The word 'objects' refers to points, lines or polygons. Similarity measures are used to obtain some degree of comparison between instances [**rosen1985match**]. They represent the criteria on which a match is based [**li2012automatically**]. The most intuitive similarity measure is the one based on position, such that we assume that two objects are homologous, when they are close in terms of distance. The most used is the euclidean distance [**beeri2004object**, **samal2004feature**, **volz2006iterative**].

The proposed approach by [**olteanu2008fusion**] is concerned with the matching of road networks (linear data) and reliefs (point data) using the belief theory. We find this approach relevant to our case since its mass modeling allows for quantification: the complete knowledge, the incomplete knowledge and the ignorance. The matching process uses two databases BD1 (reference data base) and BD2 (comparison data base), every object *obj* in BD1 is examined with every object in BD2 to find its homologous. The first step corresponds to the selection of candidates, which consists in finding, for each object in BD1, potential homologous objects in BD2, called matching candidates and noted $C_{i,(i=1,2,...,N)}$. Then, every of the matching candidates is analyzed to determine the correspondence relation. The initialization of the belief masses (second step of the process) consists, for every matching measure, in pronouncing on every of the candidates by assigning a belief to the assumptions defined for each candidate. The matching measures can be for example the Euclidean distance, the toponym-based distance or the orientation. Then, and after fusing the matching measures for each candidate (third step of the process), the fusing of the candidates is performed to have a global view of the beliefs assigned to all candidates (fourth step of the process). Finally, the decision step consists in selecting the best candidate.

3 Data imperfections in a wastewater network database

3.1 Wastewater network graph

In a wastewater network database, the nodes actually represent structures, equipment, and repairs (manholes, gullies, etc.), and the edges illustrate pipes (collection, transportation, etc.) as shown in Figure 1.

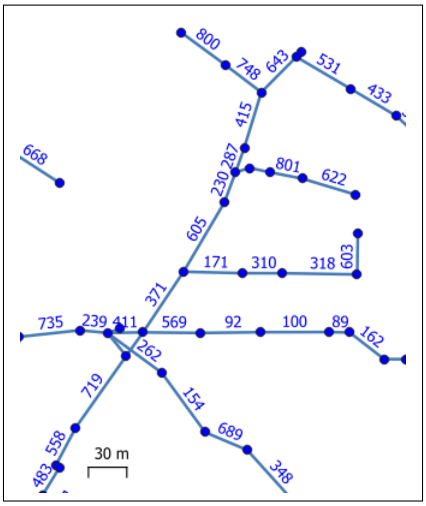


Fig. 1: Node-Edge representation of a wastewater network

3.2 Categorization of imperfections in a wastewater networks

Imperfections in data fusion are considered according to three aspects: i) imprecision, which concerns a difficulty in declaring information, ii) incompleteness, which corresponds to the absence of information and iii) incertitude, which refers to the veracity of an information: the information can be precise and complete but false [**olteanu2008fusion**, **lefevre2002belief**]. With the aim of handling these imperfections in a generic way, we propose to deepen this analysis and categorize them according to the following dimensions, relevant for the wastewater network data:

- *Nature of the object target of the matching*: node or edge.
- *Cardinality*: i) when the object is a node, it represents the number of candidates selected for an object in the comparison database that can be matched (i.e. inside the buffer around the object in the reference database), and ii) the number of candidates, among the selected candidates, that represent the real correspondent, when the object is an edge. The possible cardinalities are: 1:1 and 1:n for both nodes and edges.
- *Offset*: characterizes the shifting between the reference and the comparison objects, distance for the nodes and angle for the edges. For the edges, When the angle is equal to, or near 0, the offset is uniform. Otherwise, the offset is non-uniform.

Our categorization of wastewater networks data imperfections is illustrated in figure 2. Examples of each category are shown in figures 3, 4 and 5.

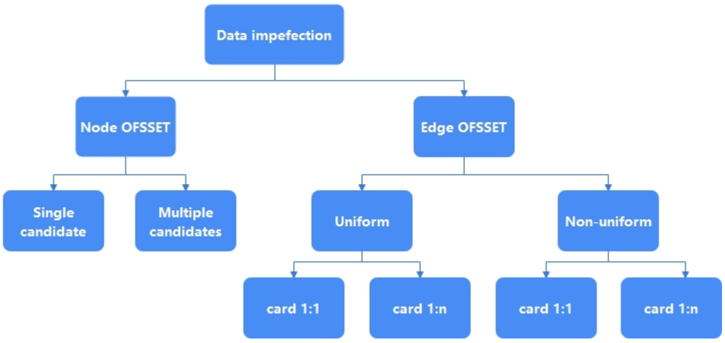
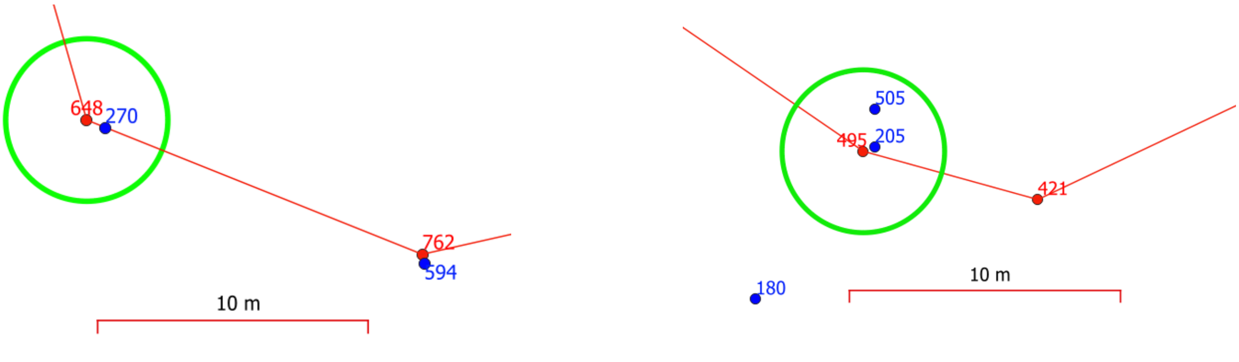


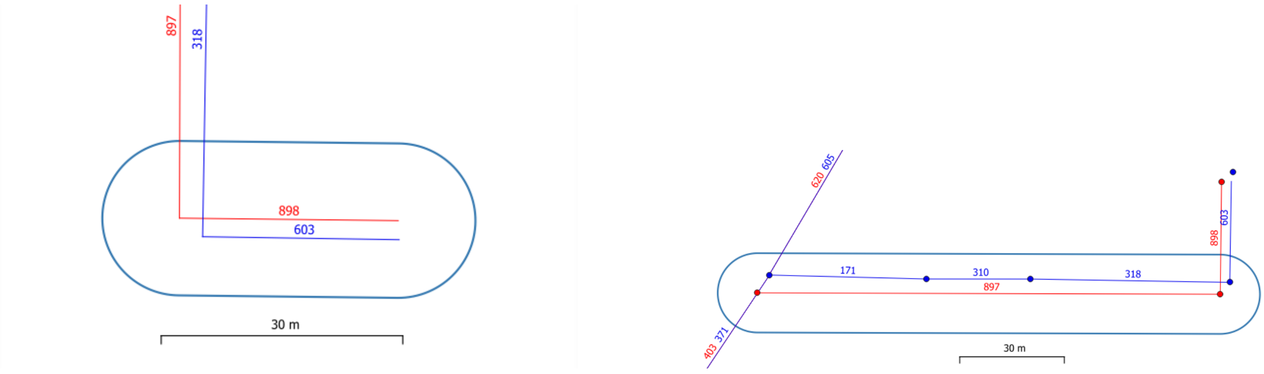
Fig. 2: Imperfection categorization of wastewater network data



(a) Single candidate

(b) Multiple candidates

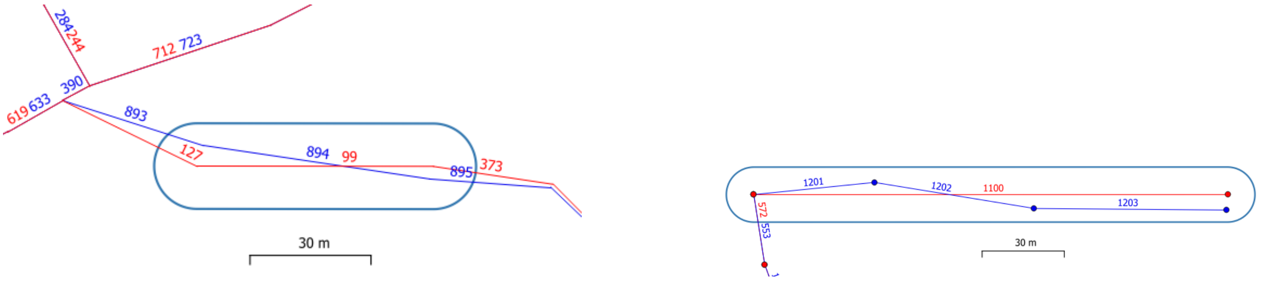
Fig. 3: Node offset



(a) Cardinality 1:1

(b) Cardinality 1:n

Fig. 4: Uniform offset of an edge



(a) Cardinality 1:1

(b) Cardinality 1:n

Fig. 5: Non-uniform offset of an edge

4 The belief theory in wastewater object matching

Our objective is the fusion of two wastewater network databases, gathered from different sources: a reference database (to be completed) and a comparison database. Inspired by the work of [**olteanu2008fusion**], we define our general frame of discernment as $\Theta = \{C_1, C_2, .., C_i, .., C_N, \text{NA}\}$, where N is the number of candidates, C_i is the confirmation that the homologous of the reference object is the candidate i and NA is the hypothesis that there is no match in the comparison database. The initialization of the belief masses is performed based on the model proposed by Appriou [**appriou1991probabilities**], where each source focuses on one hypothesis of the frame of discernment. In other words, the source of information analyzes a given hypothesis C_i, and pronounces in favor of it (C_i), against it ($\neg C_i$) or does not pronounce on it (Θ). Assignment of values to the masses functions corresponds to the semantic modeling of knowledge related to the similarity measures we adopt. Table 1, shows the similarity measures we used per object type.

	Measures of similarity			
	Euclidean distance	Hausdorff distance	Angle	Length
Node	x			
Edge		x	x	x

Table 1: The measures used for each object type

The choice of these similarity measures is relevant to the nature of our considered types of objects. Indeed, the Euclidean distance is used to measure the offset between nodes. For the edges, we use the length and the Hausdorff distance that measures the distance between two subsets of a metric space. We use the angle to measure the deviation between edges.

We have introduced several thresholds in order to transfer these similarity measures properly into masses. We give in Figure 6 the example of masses functions for the Angle measure. The same reasoning apply for the other measures.

Each edge is characterized by two endpoints. To calculate the angle between the reference edge and the candidate edge, we use the scalar product between two vectors:

$$\overrightarrow{AB} \cdot \overrightarrow{CD} = AB * BC * \cos(\theta) \tag{9}$$

The direction of the edges is not important in our case. But the formula 9 of the scalar product takes into account the directions (the values of $\cos(\Theta)$ are between -1 and 1). We propose to transform the negative values

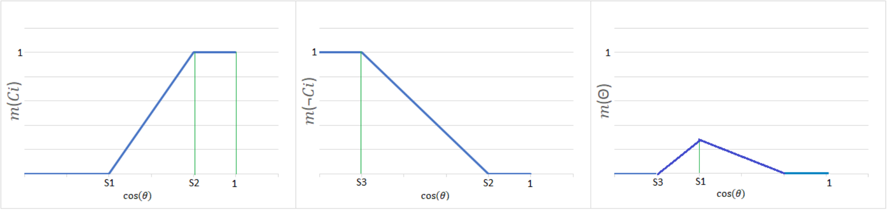


Fig. 6: Representation of knowledge for the Angle measure

into positive values to ignore the directions and then, after this step, the values of $cos(\Theta)$ are between 0 and 1. This means that the final values of the angles belong to the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

if $cos(\Theta)$ is between $S2$ and 1, we assume that the two edges are almost on each other, so the masses are distributed as follows: $m(C_i) = 1$, $\neg m(C_i) = 0$ and $m(\Theta) = 0$. Between $S1$ and $S2$, when the angle between the reference object and the comparison object is smaller, the probability that the comparison object is the homologous object is higher. The ignorance is important when the angle is in the neighborhood of $S1$, i.e. when the angle between the reference object and the comparison object is neither large enough to conclude with certainty that it is not him, nor small enough to conclude that it is the true homologue.

5 Experiments and results

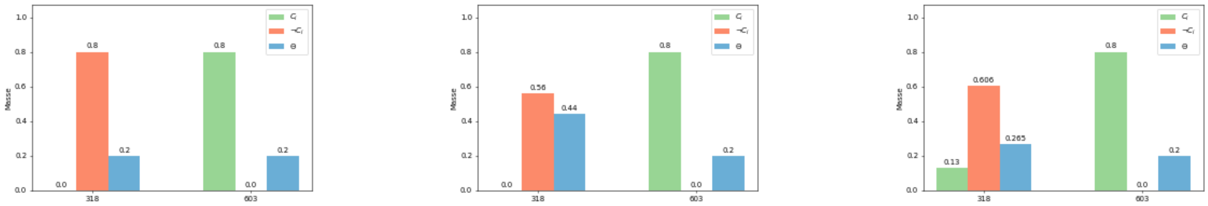
In this section, we present our experiments and results obtained for some categories of imperfections. The data set we have used are:

1. *Montpellier Méditerranée Métropole (3M)*: The official source of the wastewater networks of the city of Monpellier (south of France) containing nodes and edges. We use it as a reference database, the objective is then to complete it. All red objects in Figures 3-5 represent the nodes and edges of this database.
2. *Experimental database Ex-DB*: This is an experimental nodes and edges database that we have created to be able to manipulate all data imperfections use cases. All blue objects in Figures 3-5 represent the nodes and edges of this database.

We detail hereafter the results related to the *"Non-uniform offset of cardinality 1:1"* imperfection illustrated in Figure 5a. We used a distance buffer equal to 10 meters to select the candidates of edge number 898. Therefore, the frame for discernment is defined as follows:

$$\Theta = \{603, 318, \mathrm{NA}\} \tag{10}$$

Figure 7 shows the sets of masses for each candidate attributed for the measures Angle, Hausdorff distance and Length.



(a) Angle

(b) Hausdorff distance

(c) Length

Fig. 7: Initialisation of masses for each measure

After combining the measures (see Figure 8), we can affirm with 99% certainty that the homologous of edge 898 is 603. Furthermore, we notice that the candidate number 318 is not the homologous with 85% of certainty and a conflict of 0.12%.

After the fusion of candidates with normalisation of masses using the Dempster operator, we notice on Figure 9 that the hypothesis related to candidate 603 is clearly further away compared to the other hypotheses, with a belief mass equal to 99.2%.

When the calculation of the pignistic probability is completed (see Figure 10), the 603 hypothesis is chosen, since the pignistic probability has reached a maximum value, P(603)=0.0996. **Therefore, reference edge number 898 is matched to candidate edge number 603**.

We conducted the same experiments for the six categories of imperfection. The results are summarized in table 2.

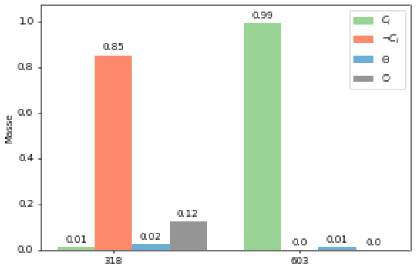


Fig. 8: Set of masses after the combination of measures

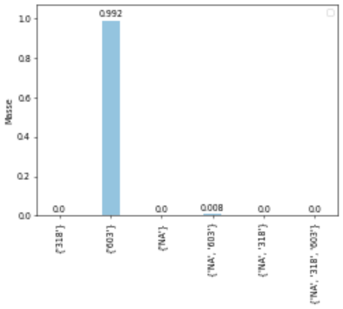


Fig. 9: Fusion of the candidates with normalization

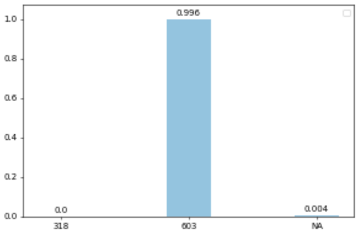


Fig. 10: Pignistic probability for each hypothesis

Imperfection category	Reference object	Frame of discernment $\Theta = \{H_1, \ldots, H_K, \text{NA}\}$	Pignistic probability							Selected hypothese
			H_1	H_2	H_3	H_4	H_5	H_6	H_7	
Node offset with a single candidate figure 3a	Node 648	$\{270, \text{NA}\}$	**0.589**	0.411	-	-	-	-	-	Node 270
Node Offset with multiple candidates figure 3b	Node 495	$\{205, 505, \text{NA}\}$	**0.659**	0.108	0.233	-	-	-	-	Node 205
Uniform offset Card 1:1 figure 4a	Edge 898	$\{318, 603, \text{NA}\}$	0	**0.996**	0.04	-	-	-	-	Edge 603
Non-uniform offset Card 1:1 figure 5a	Edge 99	$\{893, 894, 895, \text{NA}\}$	0.007	**0.984**	0.005	0.004	-	-	-	Edge 894
Uniform offset Card 1:n figure 4b	Edge 897	$\{171, 310, 318, 371, 603, 605, \text{NA}\}$	0.191	0.191	0.191	0.006	0.004	0.004	**0.414**	NA
Non-uniform offset Card 1:n figure 5b	Edge 1100	$\{553, 1201, 1202, 1203, \text{NA}\}$	0.003	0.157	0.137	0.161	**0.544**	-	-	NA

Table 2: Results for each category. The highest values are in bold.

6 Conclusions

In the context of geographic database fusion, we addressed the problem of fusing imperfect spatial data of wastewater networks. For this purpose, our contribution concerns, firstly, the categorization of the different forms of imperfections related to the nodes and edges of the wastewater networks, which allowed us to handle each category distinctly. Secondly, the application of the theory of belief in the fusion process. The results allowed us to match the reference objects in most cases. Our perspective is to further improve the results in the case of matching the edges.