Data fusion for urban network mapping : application to wastewater networks

Yassine Belghaddar

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER & L'UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH

En Géomatique & Informatique

École doctorale GAIA

Unité de recherche Hydrosciences Montpellier

En cotutelle avec Laboratoire Systèmes Intelligents et Applications, Fes, Maroc

Data fusion for urban network mapping: application to wastewater networks.

Présentée par Yassine BELGHADDAR
le 13-12-2022

Sous la direction de Ahlame BEGDOURI
et Carole DELENNE

Devant le jury composé de

[Aicha MAJDA, Professeur, UMI]	[Examinatrice]
[Laure BERTI-EQUILLE, Directrice de Recherche , IRD]	[Examinatrice]
[Azeddine ZAHI, Professeur, USMBA]	[Examinateur]
[Célia DA COSTA PEREIRA, Maître de Conférences, Univ Côte d'Azur]	[Rapporteur]
[Karim TABIA, Maître de Conférences, Univ Artois]	[Rapporteur]
[Ahlame BEGDOURI, Professeur, USMBA]	[Direction]
[Carole DELENNE, Maître de Conférences, UM]	[Direction]
[Nanée CHAHINIAN, Chargée de recherche, IRD]	[Encadrante]
[Abderrahmane SERIAI, Chargé de recherche, Berger-Levrault]	[Encadrant]

UNIVERSITÉ DE MONTPELLIER

جامعة سيدي محمد بن عبد الله – فاس
Université Sidi Mohamed Ben Abdellah - Fès

Berger Levrault

Acknowledgements

The completion of this Ph.D. would not have been possible without the guidance and support of my advisors, Drs Ahlame Begdouri, Carole Delenne, Nanée Chahinian and Abderrahmane Seriai. Their vast wisdom, wealth of experience and kindness have inspired me throughout these years. From the bottom of my heart I would like to say big thank you.

I would like to thank Drs Aicha Majda, Azeddine Zahi, Célia Da Costa Pereira, Karim Tabia and Laure Berti-Equille the members of my Ph.D. committe. I really appreciate your comments and suggestions.

My gratitude goes to my teachers, those who encouraged me during all these years of studies. A special thanks to the department of informatics in the Faculty of Science and Technology in Fez for providing the students with such a wonderful learning environment.

I must also thank CNRST Morocco, ANRT France and the company Berger-Levrault for making this multidisciplinary Ph.D. possible via the CIFRE France-Morocco program. I would like to extend my sincere thanks to the Inria team Lemon for welcoming me and counting me among their members.

I am grateful to my parent for their support during all these years. I would like to particularly express my deepest gratitude to my aunts Kabira and Hassania for all the sacrifices that have made in order for me to pursue my dreams. A special thanks to my brother Omar who've been always supportive.

To my dear friends, thanks for being the shoulder that I can always lean on. I am grateful for the countless moments of joy and happiness that we shared.

Abstract

Anglais

There are many reasons that make data management of underground networks essential: reducing the cost of repairs and expansion, running hydraulic simulations, preserving the environment etc. The available data related to these networks and more specifically wastewater ones are various, and come in different types (texts, images, GIS ,etc.) and formats (analog, digital). In addition, these multisource/multi-format data are usually incomplete, uncertain, imprecise and sometimes contradictory. Consequently, in order to extract relevant information from these inputs, a data fusion process is necessary. In fact, attributes (depth, diameter of a pipe, etc.) are always associated to a spatial representation of the objects (pipes, outfall, etc.). In this context, this work will concentrate first on using, adapting and putting forward data fusion and integration techniques to combine data collected from different sources. The second part of this thesis will be dedicated to impute and estimate missing data of a wastewater network. The result of this multidisciplinary work will be to put forward methods for fusing data and imputing the missing ones, which enables the mapping and the modelling of urban wastewater networks.

Français

Nombreuses sont les raisons qui rendent la maitrise des données des réseaux souterrains essentielle : réduire le coût des réparations et des interventions, lancer des simulations hydrauliques, préserver l'environnement etc. Les données disponibles relatives à ces réseaux et plus spécifiquement ceux d'assainissement sont diverses en terme de types (textes, images, SIG, etc.) et de formats (analogique, numérique). De plus, ces données émanant de sources multiples sont généralement incomplètes, imprécises, incertaines et parfois contradictoires. De ce fait, dans le but d'extraire l'information pertinente à partir de ces données multi-sources/multi-formats, un processus de fusion de données est nécessaire. En effet, les attributs (profondeur, diamètre d'une conduite, etc.) sont toujours associés à une représentation spatiale des objets (conduite, exutoire, etc.). Dans ce cadre, les travaux de cette thèse auront comme premier objectif d'utiliser, adapter et proposer des techniques de fusion et d'intégration de données spatiales pour combiner les données collectées à partir de plusieurs sources. En deuxième lieu, le focus sera mis sur la complétion et l'estimation des données manquantes. Le résultat de ce travail pluridisciplinaire sera la mise en place de méthodes de fusion et de complétion de données manquantes permettant la cartographie et la modélisation hydraulique d'un réseau d'assainissement urbain.

Contents

List of Figures

List of Tables

General introduction

Context

The concentration of population in a small area due to economical, agricultural or cultural factors has raised various challenges for leaders. Throughout history, the challenges evolved from providing essential requirements to the survival of the populations, such as security and food, to comfort demands like water distribution, sewerage systems and safe transport facilities (Broere, 2016). To meet the needs of this growing concentration, around the globe, underground networks are used to provide an important part of the daily services for the populations: phone lines, electricity, water, gas, etc.

The choice of burying the networks that carry these services underground is mainly to minimize accidental or intentional damage to the infrastructures and to ensure the safety of the citizens, especially from high risk networks such as gas and electrical equipment (Al-Bayati & Panzer, 2019). Today, more than half of the world population lives in cities and in 2050 this percentage is estimated to reach 68% (United Nations and Department of Economic and Social Affairs and Population Division, 2019). This rapid growth of urbanization creates more pressure for the decision-makers to satisfy the increasing need for infrastructures. Beyond the safety concerns, with a limited surface space, underground space can help cities meet these increasing demands while remaining compact (Broere, 2016). Various examples witness the intensive use of the underground space: France counts more than 2.7 million kilometers of buried and underwater infrastructure ("INERIS", 2022). In the USA, the drinking water infrastructure system alone is composed of 3.5 millions kilometers of pipes, most of which are underground (American Society of Civil Engineers, 2021).

Being buried makes the inspection and the detection of damages difficult on these utilities. Although underground infrastructures are mandatory for our modern societies, even in the most developed countries, several indicators demonstrate the need for more effort to live up to the expectations of the citizens. On average, 12 deaths and 60 injuries annually are caused by gas pipeline incidents in the USA (Congressional Research Service, 2022). In Canada, the total number of damages reported in 2019 approaches 12 000, with a societal cost estimated to be over 1.2 Billion dollars (Canadian Common Ground Alliance, 2019). The Netherlands counted 41,169 excavation damages in 2018, with a direct cost of 34.5 millions Euros (Geoff & Sakura, 2020). In the UK (Goodwin, 2005), utilities' street work have a direct impact on traf-

fic congestion, and are estimated to cost 7 billion £ annually (McMahon et al., 2005). In France every year 100,000 network damages occur during intervention (L'institut national de recherche et de sécurité (INRS), 2014).

When a network is set up, the operators record the geo-position and the various information related to the buried equipment. These data are mandatory for the managers to repair, replace and expand these networks. However, the data necessary for the operations are not always available and updated, particularly for old installations. Indeed, throughout their years of service, multiple actors may participate in the evolution of a network, sometimes even simultaneously. Due to the lack of legislation and efforts to enforce the laws, each operator is free to choose the tools and the structure to report his interventions. The updated data on the network are generally stored in several formats and representations such as images, text documents describing the activities, SIG/shapefiles, csv files, etc. As a consequence, given the networks lifespan of up to several decades, the history of operations may get lost between two successive contractors or due to obsolete tools.

Multiple studies (Al-Bayati & Panzer, 2019; Geoff & Sakura, 2020; Koschmann et al., 2021; USAG Data and Reporting Working Group, 2019) have identified inaccurate and missing information about the location of the buried pipes and cables as the primary cause of the numerous challenges facing underground networks. It is concluded in (USAG Data and Reporting Working Group, 2019), that accurate plans and more robust location and survey practices are necessary to reduce the underground utilities' damages in the UK. It was also the outcome of a case study (Koschmann et al., 2021) in Melbourne (Australia) that inaccurate mapping of underground utilities causes a significant number of damages. These studies, among others, suggest that accurate and complete knowledge about the buried utilities would reduce the incidents, thus their impacts drastically. Indeed, the impacts of the incidents are undeniably causing budget overruns and affecting our communities via traffic congestion, service interruption or compromising the safety of the individuals, especially the contractors working in the field (USAG Data and Reporting Working Group, 2019).

In addition to these issues, wastewater networks that we rely on to transport wastewater daily into a treatment plant, can damage the water environment and create public health crisis when they are inefficiently managed (Department for Environment, Food & Rural Affairs, 2002). To avoid such events, mainly due to leaking pipes, the operators rely on various tactics such as field inspection, the type of pipe material, history of interventions, age of the network or reports from inhabitants. Beside these traditional strategies, researchers use hydrodynamic modelling to simulate water flow under different conditions. This allows them to compare the actual network

behaviour to the simulations, hence to diagnose anomalies and draw different hypothesis about their sources but also to identify and anticipate potential leaks and deficiencies. Still, missing data on the networks makes this task inefficient and inaccurate. In the context of climate change, it is now urgent than ever, to protect water resources from any threat. Accordingly, in the last decade, important efforts have been deployed to improve the management of underground networks and particularly sewerage networks around the globe. In the UK, a major initiative entitled Mapping the Underworld ("Mapping the Underworld", 2022; Muggleton & Rustighi, 2013) started in 2005 to develop means to locate, map and record the position of all buried utilities. As for water infrastructures, the authorities through national and European regulations impose several protocols and criteria to limit and reduce the risks of environmental degradation (Department for Environment, Food & Rural Affairs, 2002). In the US, $50 billions, over the period of 2022 and 2026, which represent the single largest investment in clean water by the USA federal government (Bipartisan Infrastructure Law: State Revolving Funds Implementation Memorandum, 2022), are granted to the Environmental Protection Agency (EPA) to strengthen drinking water and wastewater systems.

To address the issues related to inaccurate and missing data of underground networks, in addition to manual inspection such as manholes observations, operators around the globe use various methods to estimate the locations of the network's objects. Techniques based on signal transmission and electromagnetic technologies are the most popular ones and the main methods to locate buried pipes and cables. For instance, Ground Penetrating Radar (Chen & Cohn, 2011), low-frequency vibro-acoustic (Muggleton & Rustighi, 2013) or magnetic induction (Sun et al., 2011). Due to the complexity of the task and the difficulty in interpreting the sensor data, these techniques are often guided by/combined to, the available information about the targeted network (Chen & Cohn, 2011; Hafsi et al., 2017), or require to be paired with other approaches (Muggleton & Rustighi, 2013). Recently, given the remarkable progress in image processing using machine learning, visible objects on the ground such as manhole covers and stormwater inlets are located using images (Boller et al., 2019; Commandre et al., 2017). Also, data mining techniques are applied to extract accessible knowledge on a network from the web (Chahinian et al., 2021). Although this progress is important, still none of these approaches is sufficient to produce accurate maps of the networks. However, they can be combined to produce maps with less imperfections. Besides, since each method usually focuses on limited components of the network, and due to the disparate structures and formats of the recorded data, a generic framework in which all these propositions could be integrated is necessary to help aggregate and combine all the network data.

Data on underground networks, particularly wastewater networks, could be divided into two categories. The first consists of spatial data, usually represented by a graph, which indicates the position of objects such as manholes and pipes and their relationships. The second represents the properties related to each object, usually alpha-numerical data organized in the form of tables, such as pipes' diameters or the type of materials. Although sensor and image based propositions offer innovative methods to collect data, the attributes of the network's objects cannot be collected using images, and only with important resources: multiple sensors and prior knowledge (Bilal et al., 2018; Hafsi et al., 2017), attributes such as diameter and type of material, can be identified by sensors. Thus, alternatives solutions to address missing attribute values must be investigated.

Based on the outlined challenges regarding the management of underground networks particularly wastewater networks, we introduce, in the following paragraph, the scientific questions that we addressed in this thesis.

Scientific challenges

This PhD research started with the aim to improve knowledge about wastewater networks. In addition to the conclusions reached based on the literature review, our interaction with network managers and the study of publicly available data lead us to identify four main aspects that must be considered in order to achieve our goal:

- Network data can be collected from multiple sources.
- Network data can be stored in different structures and formats.
- Network spatial data provided by a single source are often imperfect.
- Network attributes suffer mainly from missing values.

The diversity of the sources' structures and formats constraint us to propose a generic solution in which all the data sources are exploited in order to achieve more accurate an complete wastewater databases. As for imperfections, from which suffer often all the sources individually, there is no rigid definition. According to 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 (Dubois & Prade, 2009; Smets, 1997). For wastewater networks we distinguish between three types of imperfections:

- Incompleteness: indicate missing information.
- Imprecision: the doubt about multiple possible values.

- Uncertainty: related to the veracity of an assertion.

The collection of heterogeneous and multi-source information inevitably raises the problem of fusing all this data. Therefore, two main questions were derived from these aspects:

- How to fuse or combine data obtained from multiple sources in the context of a wastewater network?
- What solutions could be put forward for wastewater network missing data?

Data fusion

In the last three decades, significant efforts have been deployed to address the problem of fusing or combining data collected from different sources. Numerous definitions of data fusion were proposed. A popular one (Hall & Llinas, 1997), suggests that *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.* In (Wald, 1999), *data fusion is a formal framework in which are expressed means and tools for the alliance of data originating from different sources. It aims at obtaining information of greater quality; the exact definition of 'greater quality' will depend upon the application.* Although the definitions vary between studies, in general, almost all of them agree that data fusion is a process to achieve better results compared to those obtained with a single data source. This is particularly true for wastewater networks, given that single data sources often provide imperfect data. The applications of data fusion are various. First, the focus was on the military domain, such as target tracking (D. Smith & Singh, 2006), then it extended to many others, such as image processing (Simone et al., 2002) and transportation systems (El Faouzi et al., 2011).

A large number of studies proposed a conceptualization of fusion systems as a base to conduct data fusion operations. The Joint Directors of Laboratories (JDL) model (White, 1987) and Dasarathy's framework (Dasarathy, 1997) are among the popular ones. The JDL model, oriented for military application, divided the fusion process into 5 different levels of abstraction: source preprocessing, object, situation, threat and process refinement. Dasarathy's framework structured the levels of fusion depending on the input and the output: data-data, data-feature, feature-feature, feature-decision and decision-decision. Nevertheless, these models are more theoretical than used in practice, where often data fusion architecture is defined following the judgment of the scientist(s) conducting the fusion operation based on the available data and the task at hand (Raol, 2015). In addition, data fusion systems can be classified based on other aspects such as the type of architecture (centralized, distributed

or hybrid) and the relation between the data sources (complementary, redundant or cooperative). A detailed classification can be found in (Raol, 2015) and (Castanedo, 2013). Thus, the first question that we address in this manuscript is the following:

In view of data fusion operations, how can wastewater data sources be modeled given their heterogeneous and imperfect nature?

Although there are several factors that make data fusion a difficult operation like data heterogeneity, desired outputs and time constraints, data imperfection is the most fundamental challenging problem of data fusion systems (Khaleghi et al., 2013). Regardless of the fusion model (JDL, Dasarathy or others), a processing technique capable of handling the imperfections is mandatory. Accordingly, an important part of the efforts on data fusion systems were dedicated to this issue.

Techniques based on a set of mathematical theories were used and developed to deal with imperfections as part of data fusion systems. Each mathematical theory is designed to deal with specific types of imperfections. Therefore, the quality of a particular data processing technique is related to its ability to handle different types of imperfections (Appriou, 2014). For instance, the probability theory, the oldest and the most widely used approach, models uncertainty. The fuzzy set theory is intended to represent imprecision or vagueness and the Dempster-Shafer theory is considered more powerful due to its ability to represent both uncertainty and imprecision.

The spatial position of the network's components is the central and essential information in wastewater network data, considering the attributes are always associated to a spatial representation of the objects. Therefore, a desirable data fusion process for wastewater network data must focus on the spatial dimension.

The operation of combining spatial data is refereed to as data integration or map 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 (Devogele et al., 1998). Data conflation is defined as the process of creating a new dataset based on multiple datasets that cover the same spatial area (Tong et al., 2009). Whether the goal is to create a schema to query input instances or to create a new dataset from available sources, object matching, which aims at identifying objects that represent the same real object, is identified as the most important, difficult and challenging step in both data integration and conflation (Costes & Perret, 2019; Song et al., 2011; Volz, 2006). Hence, the second and main scientific problem that we address in this thesis is the following:

How can object matching of wastewater networks be achieved and how can the imperfections be modelled?

Data imputation

Whether in medicine (Bell et al., 2014), transportation (B. L. Smith et al., 2003), hydrology (Aissia et al., 2017), missing data is a common research problem in almost all domains that deal with data. This issue is addressed for different purposes, often it is to avoid inaccurate conclusions about a subject or phenomena. For wastewater networks it is to efficiently manage these infrastructures.

The mechanisms that lead to missing data are diverse. The classification adopted by the research community is the one proposed by the authors of (R. J. Little & Rubin, 2019; Rubin, 1976), which divide these mechanisms into three categories:

- Missing completely at random (MCAR): when missing values are independent of the variable under investigation and the observed data.
- Missing at random (MAR): when missing values depends on the observed data.
- Not missing at random (NMAR): when missing values depends on the variable under investigation and the observed data.

These assumptions about the categories of missing data are important for choosing or proposing an appropriate method to deal with the data at hand. For instance, using the mean metric to replace the missing values is relatively more appropriate for MCAR cases than MAR or NMAR ones. For wastewater network data, the values are MCAR since they are due to unreported interventions, lost data between different operators or because of storage technical issues.

Missing data can be handled simply by ignoring or deleting the missing instances (T. Little et al., 2013). First, for wastewater networks, this is not an option since data are required for their management, contrary to other domains where the goal is to analyse a phenomena such as the occurrence of extreme event in hydrology (Hamzah et al., 2020), or the interpretation of clinical data (Pedersen et al., 2017). Second, this leads to various problems such as biased results, failures in predictions, decreased information content and unbalanced datasets (Farhangfar et al., 2007; Young et al., 2011). Therefore, under the name of missing value estimation and missing value imputation techniques, multiple methods have been proposed to replace or substitute the missing values with estimated or predicted values.

A large panoply of imputation techniques have been developed and can be classified according to different criteria (Emmanuel et al., 2021; Hamzah et al., 2020; Young et al., 2011) some are based on the statistical properties of the datasets, which are widely used in almost all domains and applications. For instance, authors in (Jerez et al., 2010) used the mean and the mode to impute missing values of a breast cancer database. The expectation maximization algorithm, which estimates the parameters of a probability distribution from incomplete data by maximizing the likelihood of the available data iteratively (Dempster et al., 1977; Schneider, 2001), is applied in (Nelwamondo et al., 2007) to HIV and power plant databases.

In the last decade, a new branch of missing imputation techniques based on Machine Learning (ML) models were proposed. The particularity of ML techniques, compared to traditional methods, is the ability to directly learn patterns and models from the data without explicitly requiring complex hand crafted parameters or restricting the models to limited hypothesis. The learned patterns are then used to predict the missing values. K-Nearest Neighbor and artificial neural networks are among the most applied methods in order to deal with missing data (Alabadla et al., 2022; Emmanuel et al., 2021; Hamzah et al., 2020; Huang et al., 2017; Nishanth & Ravi, 2016).

ML techniques, particularly artificial neural networks, have already proven their efficiency in various domains such as image processing (Krizhevsky et al., 2012) and natural language processing (Mikolov et al., 2013). For imputation applications, ML models have also outperformed traditional methods in many cases (Hamzah et al., 2020; Jadhav et al., 2019; Xu et al., 2020). However, no method can be considered superior and the performances can vary between datasets (Osman et al., 2018; Young et al., 2011). Nevertheless, for wastewater networks the rate of missing values is high and a systematic review (Alabadla et al., 2022), regarding the use of ML techniques in imputation, indicates that ML approaches can deal with high missingness rates while maintaining a small error regardless of the dataset size. Hence, further investigation of ML techniques should be considered to address the issue of missing values in wastewater network databases.

A more recent type of neural network called Graph Neural Network (GNN) have been proposed in the literature to learn patterns directly from graph structures (Scarselli et al., 2008; Zhou et al., 2019). Besides using the features, GNN exploit the structural relationships between the components of a graph (nodes and edges) to learn hidden patterns. GNNs have gained a lot of attention in last few years and have been successful in multiple applications such as social recommendation (Fan et al., 2019), chemistry (Duvenaud et al., 2015) and transportation (Guo et al., 2019). Given that

a wastewater network can be modeled as a graph, the third scientific question that we address in this thesis is the following:

How can the missing attribute values of wastewater networks be imputed and how can the structure of the graph be used for this purpose?

Imputation techniques should be selected based on the target domain and its characteristics rather than only previous performances (Alabadla et al., 2022). Indeed, another way to increase the accuracy of the imputed values is to consider external information and domain knowledge (Armina et al., 2017). This has been successfully applied in different domains. For instance in biology (Gan et al., 2006), where the authors exploited the biological constraints to reduce imputation errors for micoarrays gene data. In medicine (Kamkhad et al., 2020), the semantic relationships between the attributes (time, localisation, number of population, etc.) were exploited in the imputation process to predict the missing values. Hence, this lead us to the fourth scientific research question that we address in this manuscript:

What domain knowledge or external information could be useful for both data fusion and imputation techniques and how can they be used?

The scientific questions above show that this PhD research topic is multidisciplinary. In fact, to tackle these questions four domains must be investigated. First, a profound understanding of the application domain, which is wastewater networks and their daily management, is required. Second, to encapsulate all the contributions in one product or entity due to the industrial context, knowledge about software engineering and data modeling is needed. Third, data fusion and object matching techniques must be studied to achieve the desired integration of the data sources. Fourth, Missing Value Imputation skills are necessary to address the problem of data incompleteness.

Contributions and thesis structure

To address the scientific questions that we outlined, we conducted a study of the concerned domains: wastewater management, data modelling, data fusion/object matching and Missing Value Imputation (MVI), then we proposed the following contributions:

1. Meta-modelling: as a base and an abstraction for data fusion operations, we proposed a generic meta-model of wastewater network data sources, with the aim of conducting data fusion operations. Our meta-model supports

imperfection modelling at data source level as well as at network object position and attribute levels, allowing thus formal fusion operations to be conducted efficiently and reliably. Chapter 1 of this document is dedicated to this contribution.

2. Object matching and data fusion: to integrate wastewater network data collected from different sources, we proposed a line matching approach where similarity measures are combined using Deampster-Shafer's theory. This contribution is detailed in Chapter 2.

3. Missing value imputation: this contribution can be divided into three sub-contributions:
   - Based on domain knowledge and external information, we proposed a set of algorithms to estimate required attributes to run a hydraulic simulation.
   - In addition to knowledge domain, we used the network's topology through Graph Neural Networks to impute missing values of the network's attributes.
   - We proposed a novel GNN to map wastewater networks from the manholes positions.

   The details of this contribution is on Chapter 3.

Publications

Journal Papers

- Belghaddar, Y., Chahinian, N., Seriai, A., Begdouri, A., Abdou, R., & Delenne, C. (2021). Graph convolutional networks: Application to database completion of wastewater networks. Water, 13(12), 1681.
- Belghaddar, Y., Seriai, A., Begdouri, A., Delenne, C., Chahinian, N., Rima, B., & Derras, M. (2022). Towards a generic fusion framework for underground networks involving model-driven engineering domain. International Journal of Information Science and Technology, 6(2), 8-19.
- Belghaddar, Y., Delenne, C., Chahinian, N., Seriai, A., Et-targuy, O., & Begdouri, A. (2022). Line matching relying on the DS-theory: application to wastewater networks. International Journal of Approximate Reasoning (In Review).

Conference Papers

- Bel-Ghaddar, Y., Seriai, A., Begdouri, A., Delenne, C., Chahinian, N., & Derras, M. (2021, June). Combining model-driven engineering and sewerage networks: towards a generic representation. In 2020 6th IEEE Congress on Information Science and Technology (CiSt) (pp. 48-53). IEEE.
- Belghaddar, Y., Delenne, C., Chahinian, N., Seriai, A., & Begdouri, A. (2022, July). Parametrization of a wastewater hydraulic model under incomplete data constraint. (HIC 2022).
- Et-targuy, O., Belghaddar, Y., Begdouri, A., Chahinian, N., Seriai, A., Delenne, C. (June 2022). Data imperfection categorization for wastewater object matching using the belief theory. Advanced Intelligent Systems for applied Computing Sciences.

Project context

This work was carried out within the framework of the CIFRE-France/Morocco Program. It is a collaboration between three entities:

- University of Montpellier in France.
- University Sidi Mohamed Ben Abdellah in Fez, Morocco.
- Berger-Levrault, a private company that provides software solutions to increase the performance of public and private stakeholders.

1. Meta-Modeling of wastewater networks data

The disparity of structures and formats adopted by the actors involved in the management of a wastewater network coupled to the diversity of data collection approaches such as radars and images, make the communication, exchange and monitoring of data difficult. The main goal of this PhD research topic is to propose a solution that enables the combination of all the data sources that may be used to confirm the managers' data and enhance knowledge about a given wastewater network. For this purpose, regardless of the method and the architecture used for the fusion, data extracted from the sources must be organized in similar schemes to conduct fusion operations. In this chapter, adapted from the work (Belghaddar et al., 2022), published in the International Journal of Information Science & Technology, we answer the first scientific question by proposing a generic data model for the fusion of wastewater network data. Our meta-model supports imperfection modelling at data source level as well as at network object position and attribute levels, allowing thus formal fusion operations to be conducted efficiently and reliably. To validate our meta-model, we implemented it using a data analysis and reengineering platform called Moose, and carried out a test on the town of Prades-le-Lez (France). We took into account three data sources providing information on the node positions of the wastewater network: 1- the official network map, 2- a high resolution aerial image database and 3- a Google Street View database. As shown in the results section, we were able to reliably perform data monitoring and visualization requests on real heterogeneous multi-source data related to a specific wastewater network.

1.1. Introduction

By the time a sewerage network is set up, its graph (where the nodes represent the manholes or inlet grates and the edges represent the pipes) is mapped for the first time. Later on, several actions, such as repairs or expansions, may occur in the field according to the new needs of the citizens (ASTEE, 2015; Bernold et al., 2003). These modifications to the network are recorded by the operators that have conducted the actions. The updated data on the sewerage network are generally stored in several formats and representations such as images, text documents describing the activities, SIG/shapefiles, csv files, etc. For example, data on the sewerage networks published on the French Government's open access portal ("Open platform for French public data", 2015) shows this diversity. Consequently, the combination of data from different sources and eras raises problems of consistency (data conflict), which may be due to differences in granularity or accuracy of the data sources (Chen & Cohn, 2011) and requires the establishment of a methodological framework for collecting, centralizing, updating and data archiving in order to facilitate information sharing and communication between the managers. In our vision of data fusion, we consider the uncertainties related to each type of collected information in order, for example, to anticipate and react promptly to potential dysfunctions or to quantify their impact on the results of a numerical simulation of flows in the sewerage network. In this context, and as a first step of our work, we propose a meta-model for aggregation, control and analysis of data sources related to sewerage networks, before elaborating adapted algorithms to merge heterogeneous multi-sources data. This chapter is structured as follows, Section 1.2 introduces the context of our work. We explain the motivation behind this work in Section 1.3 and we present the state of the art and related works in Section 1.4. Section 1.5 describes our meta-model and its specific viewpoints. Section 1.6 demonstrate the implementation and the results of instantiation of our meta-model in Moose (Ducasse et al., 2005), using real data of a sewerage network provided by multiple sources. Section 1.7 concludes this study.

1.2. Context

1.2.1 Sewerage networks and management challenges

Sewerage system is a network for collecting and transporting wastewater and storm water to a treatment plant, also called combined sewer system. When a network collects these two types separately, it is called separated sewer system. To set up a sewerage system and make its progress in a territory, different institutional and

operational actors are involved. The ministry in charge of sanitation or the ministry in charge of local communities, define sanitation policies and strategies as well as the regulatory framework at national level. The local authorities ensure the respect of regulations related to the quality of sanitation services. Finally, the contracting authorities (municipalities or state agencies) are responsible for the development of services, their quality and sustainability. For the implementation, monitoring and control of these services, they call for actors such as service operators, local associations, and development partners (founders, NGOs, and design offices). The sanitation supply is not only limited to infrastructure installation. Maintenance tasks such as reparation, expansion, damage anticipation and scheduling of interventions are all necessary actions to ensure a permanent and transparent services. Planning is an important task for decision making. It allows to develop a vision of needs in space and time, to quantify and prioritize them in order, among others, to direct funding towards the most necessary investments and at reasonable costs.

On a territory, infrastructures are large and must be managed in a collaborative way by the diverse involved actors. Urbanization and the concentration of populations in cities engender the increase of the dimensions of sewerage networks. For example, in France, this heritage consists of approximately 337,000 km of collectors (ASTEE, 2015). The 2012 reform “DT-DICT” (“Legifrance”, 2012), as part of the network detection process indicates that France is covered by more than four million kilometres of networks, one third of which are aerial and two thirds are buried or underwater.

The improvement of underground networks, particularly for sewerage networks, has several advantages, especially the impact on public health and environment preservation through the protection of water resources against pollution. The administrative management and the techniques of interventions play a key role in these challenges since they help to reduce damage costs induced from services interruptions and floods.

The expenses associated with the management of these infrastructures are high, particularly repairs, since the components of these networks are subject to degradation and damage caused by several factors: age, environment, etc. Furthermore, the costs of urgent and unexpected operations are far higher than the ones anticipated (ASTEE, 2015).

In this context, improving knowledge on the state of these networks, mostly underknown, becomes a priority. Indeed, digital technologies such as Geographic Information Systems (GIS) and Computerised Maintenance Management Systems (CMMS) bring great added value and are officially increasingly adopted. For example, the regulations associated to the environment code in France require stakeholders to have

digital and precise cartography for sensitive underground networks since January 1st, 2019 in urban units and from January 1st, 2026 in other cases.

These solutions are particularly useful for making spatial and geographic data available to the various actors, to facilitate their communication for optimal decision-making as well as to improve the administrative, economic, and financial management of this type of networks and of interventions to take place near these underground networks.

Since the majority of the components of sewerage networks are buried, collecting and detecting information about these infrastructures is challenging. However, with the advent of new technical solutions, different methods were used to extract information such as Ground Penetrating Radars (Hafsi et al., 2017). Nevertheless, these propositions are usually applicable and efficient under certain conditions and constraints, such as the availability of Google Street View images in the case of (Boller et al., 2019). Besides, each study usually focuses on limited components of the network. For instance, in (Boller et al., 2019; Commandre et al., 2017) only manhole positions are being collected. In (Chen & Cohn, 2011), the focus is on collecting pipes positions. As for (Kabir et al., 2020), the attributes of the objects are the target. Thus, a generic framework in which all these propositions could be integrated is necessary to help collect all the network data.

1.2.2 Sewerage network representation

A sewerage network is represented by a graph composed of nodes and edges. Nodes represent manholes, equipment, repairs, etc. 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 of the inspection areas for the objects. In recent years, storage and data management solutions for sewerage networks have evolved. Currently, most managers use Geographic Information Systems to create, edit, view, and analyze these data. The data structures and formats supported by these systems are diverse: relational databases, Shapefiles, GeoJSON and CSV files, etc. Moreover, to study the impact of some parameters, such as the discharge rate of consumers into the networks, specialists use hydraulic simulation software.

Although the applications are various, the digital representation of the data remains almost identical in the different solutions:

- Spatial data that are represented by geometric shapes and their relationships: points and lines (Figure 1).

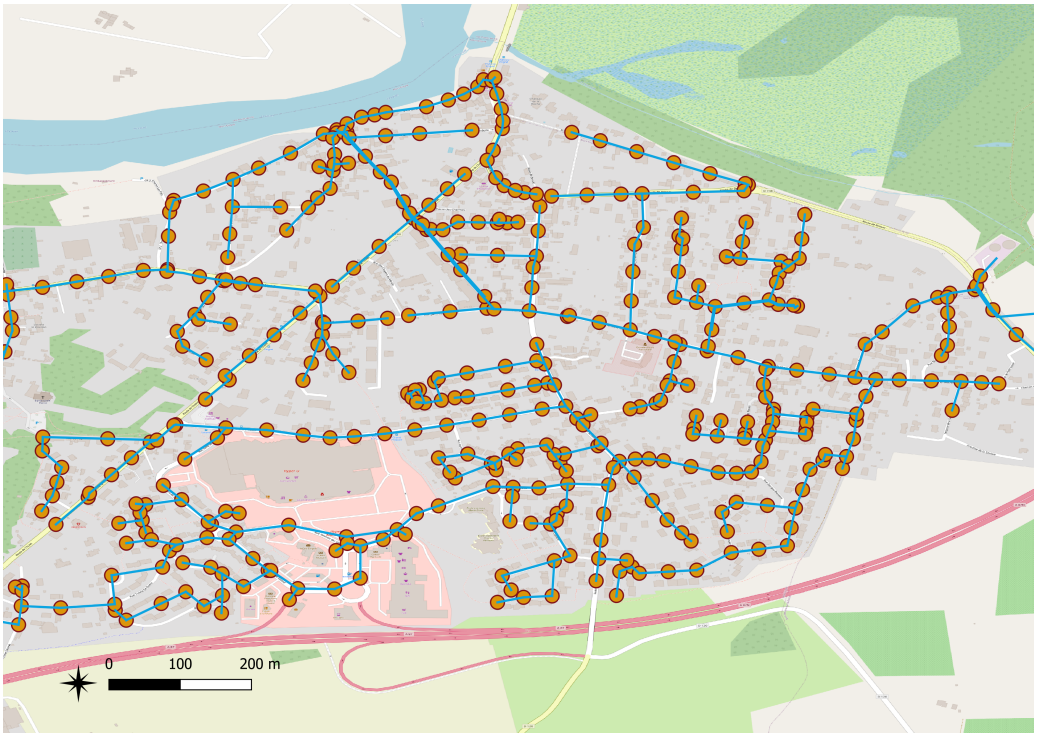


Figure 1 Example of spatial representation of sewerage networks.

– Atributes that are listed in attribute tables where each record is associated with a network object (Figure 2).

1.3. Problematic

When using GIS solutions, data processing includes acquisition, digitization, import, export, and visualization of geographical data. The acquisition can be carried out directly in the field allowing real time data collection and mapping. In addition,

	commune	ecoul	nom_voie	typer	dimensions	materiau	id
16	SAINT-JEAN-DE...	GRAVITAIRE	CAMILLE CLAU...	EAUX USEES	200	FONTE	15
17	GRABELS	GRAVITAIRE	NULL	EAUX USEES	999	INCONNU	16
18	MONTPELLIER	GRAVITAIRE	DU FAUBOURG ...	UNITAIRE	999	INCONNU	17
19	PRADES-LE-LEZ	GRAVITAIRE	NULL	EAUX USEES	200	INCONNU	18
20	LATTES	GRAVITAIRE	NULL	EAUX USEES	999	INCONNU	19
21	SAINT-JEAN-DE...	GRAVITAIRE	RUE CAMILL...	EAUX USEES	200	PVC	20
22	LE CRES	GRAVITAIRE	NULL	EAUX USEES	999	INCONNU	21
23	SAINT-JEAN-DE...	GRAVITAIRE	RUE CAMILL...	EAUX USEES	200	PVC	22

Figure 2 Example of attributes table.

advanced data processing may allow considering multi-source data, spatial analysis through interactive queries and maps overlays.

Although the use of digital maps is increasingly adopted, there are still large communities in the world where maps and geographic data are still analog, making their use and update difficult. The detection and digital mapping of buried networks by semi-automatic or automatic approaches is a real scientific and technological challenge. Therefore, there is a large conceptual, technical and semantic gap between the analog and digital mapping models.

The attributes and characteristics associated with the various objects constituting a network are not all available at a given time (Belghaddar et al., 2021; Kabir et al., 2020). This is partly explained by the fact that the networks undergo expansions and repairs but not properly tracked and reported, or through the interventions at different stages by actors, other than the operators who ensure the continuous functioning of the supply services. However, these attributes may be reported elsewhere, for example, in public databases, calls of tender, repair reports or even in press articles reporting damages.

In addition, since information and communications technology are easy to reach and use, operators currently have access to several sources from which they can collect useful data before interventions in the field, such as images, analogue maps, reports of interventions, sensors, etc. The heterogeneity of the sources makes the extraction of relevant information and its combination a complex and time-consuming task.

On the other hand, imperfections may be found in these datasets and sources, namely inconsistency (abandoned pipes which still appear on the maps), missing attribute values for some objects, uncertain and sometimes contradictory values. All of these aspects represent various obstacles to operators when merging the data.

Combining multi-source data also requires a unified data model to allow the centralization, updating, archiving, and monitoring of these data. Indeed, we have analyzed digital databases related to sewerage networks to understand the semantics of their data, their relationships and determine their differences. Since the associated data models, when they exist, are rarely available to the public, we proceeded by inferring them from data. In our study, we have used the data provided by reliable sources, particularly, the French open data repository ("Open platform for French public data", 2015). Among the suppliers are the urban community of the South-East of Toulouse Sicoval, Data Angers, and the region of Pays de la Loire.

As a result of studying these different sources, we identified the following constraints:

- The data models adopted by operators are different. Therefore, exchanging and reusing data is difficult.
- The models do not comply with computer design and modelling rules and standards.
- The attributes provided by the stakeholders are related to their fields of activity. For example, a company specialized in hydraulic modelling provides precise information on the flow of water in a pipe, while the same information is generally missing in the data provided by another entity expert in the field of infrastructure repairs.
- The history of interventions, necessary for anticipating repairs, is rarely considered in these models.

Thus, a generic model for business data and data sources is needed to overcome the conceptual and semantic gap between existing digital mapping models and to implement an optimal data fusion approach.

1.4. State of the art

1.4.1 Meta-models

Several definitions have been proposed for the term "meta-model". According to (Object Management Group, 2006), a meta-model is a model used to model modelling itself. A Meta-model is a model that defines the structure of a modelling language (Da Silva, 2015). The basic idea of a meta-model is to identify the general concepts in a given problem domain and the relations used to describe models (Gascueña et al., 2012). This generality, which we also seek to satisfy in our proposal, is one of the most important axis that meta-modelling has made, i.e. the creation of meta-models, one of the most important approaches for modelling. Instances of a meta-model are models that must satisfy the meta-model specifications. They enable target systems to be modelled in a consistent and homogeneous manner.

Monitoring activities is one of the domains where meta-models are used. For example, a meta-model for properties associated with software during execution is presented in (Bertolino et al., 2011) to ensure the quality of the software and its dynamic adaptation after deployment. Indeed, these properties provide a means to assess and improve the resilience of software through the adaptation and anticipation of abnormal situations. The instances of this meta-model are in this case a model with monitoring properties adapted to the target software. In (la Fosse et al., 2020), the authors

propose a meta-model for the monitoring of cyber-physical systems (CPS), particularly sensor and actuator networks which require valid data and good coordination between sensors and actuators for their operation.

1.4.2 Sewerage networks business modelling: related works

To help decision makers in collecting the data necessary for interventions and to diversify their data sources, some solutions have been published. For example: in (Commandre et al., 2017), the authors apply deep neural networks to detect the position of manholes, visible on the ground, from a high-resolution image. In (Chen & Cohn, 2011), to create the cartography of underground networks, researchers use Bayesian fusion techniques to combine hypotheses extracted from the Ground Penetrating Radar (GPR), the spatial location of surveyed manholes, as well as the expectations from the statutory records. However, none of these approaches have been submitted along with a data model.

The work in (Abdelbaki & Zerouali, 2012) is an attempt to design a business data model for sewerage networks to build the digital map of the sanitation network for a municipality in Algeria and to contribute to its efficient management. The authors propose, from the inventory of the various available data sources, a conceptual data model on which the necessary objects for the management and their relationships are listed.

At the initiative of the Aquitaine region and a public interest group (Planning and risk management), the Commission of Data Validation for Spatialized Information (COVADIS) has published a data standard for drinking water and sewerage networks intended for French municipalities (COVADIS, 2019). The committee presented a class diagram describing the minimum and necessary data to be used by the actors participating in the management of these networks (municipalities, PEIC[1], delegates public services, etc.) for the purpose of a simple data exchange between them.

Since this standard describes the minimum necessary, but sufficient, data for the management of the water networks, and since it is also a standard to be adopted at a nationwide level, we adopt it in our work as a business data model. We present in the following the subpart of the COVADIS class diagram related to the sewerage networks (Figure 3).

It is composed of 4 main classes: Node, Pipe, Reparation and Meta-data whose attributes and related possible values are listed:

[1] Public Establishment for Inter-municipal Cooperation.

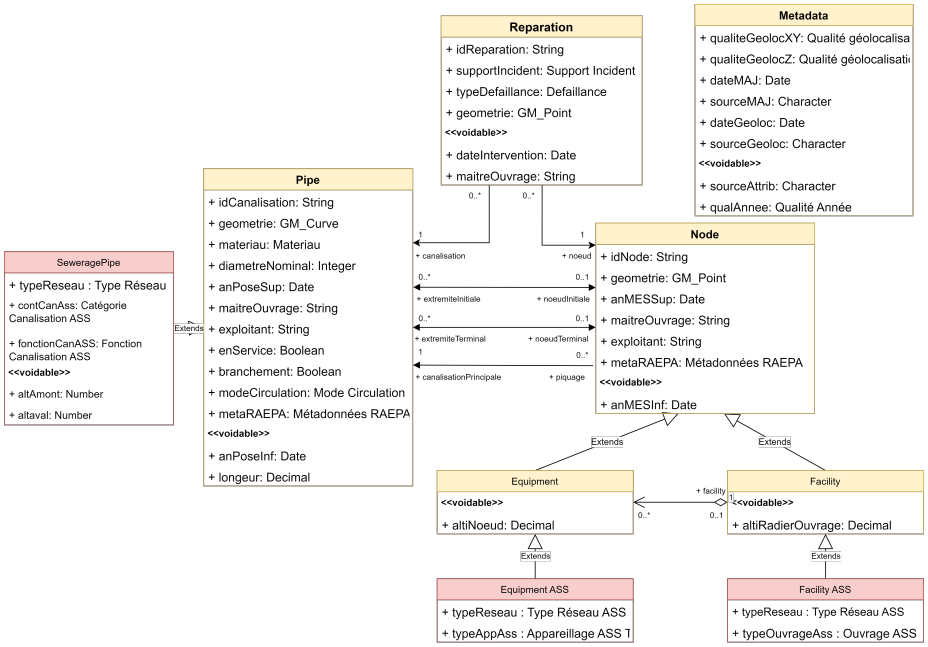


Figure 3 COVADIS sewerage networks business model (COVADIS, 2019).

- Nodes: represented geometrically by points, they illustrate apparatus (valve, counter, etc.) or manholes.
- Pipes: represented geometrically by lines, they are classified into several categories: wastewater, rainwater, etc. Each of the pipes has two end Nodes.
- Repairs: geometrically represented by points, they refer to interventions made in Nodes or Pipes.
- Metadata: are data used to qualify the information of the classes Nodes and Pipes. Namely the name of the source, the date of the last update, the reliability of the year of installation and the quality of the geolocation within respect to the 2012 decree ("Legifrance", 2012), which defines 3 precision classes: less than 40 cm, in the range 40 cm and 1.5 m and greater than 1.5 m.

1.4.3 Sewerage networks and Big Data

Nowadays, we are witnessing an intensive production of data. Every day, a huge amount of data is produced by companies, on social networks, during transactions or through sensors, that conventional computer tools can no longer process and analyse. The research work around these masses of data, also called Big Data, is a response to these obstacles.

On the other hand, and despite the small amount of available data on sewerage networks compared to the Big Data, they share two important characteristics:

- The multitude of data sources.
- The heterogeneity of the data.

The research works in the domain of the underground networks is mainly confidential (Hafsi et al., 2017). Therefore, the number of publications related to Big Data is more important compared to sewerage networks. To our knowledge, there is no data model for data sources of underground networks. To fill this gap, and since the two characters of the multitude and heterogeneity of the sources have already been examined in Big Data (see for example (Dong & Srivastava, 2013) or (Boury-Brisset, 2013)), we have chosen to draw inspiration from the solutions proposed in this field.

The available Big Data systems and platforms are not identical, as they are obtained through multiple providers whose vision is not uniform. In (Erraissi & Belangour, 2018) using model-driven engineering, the authors propose a platform-independent meta-model to describe the structures of data sources involved in feeding these large volumes of data, thus allowing programmers to create applications compatible with various products. Given that Big Data include three heterogeneous formats: Unstructured, Semi-Structured and Structured (Oussous et al., 2018), the authors propose a Meta-Modelling of the three types of data sources as follows:

- Structured: data whose set of possible values are determined and known in advance, such as relational databases.
- Semi-structured: data that have not been organized into a specialized repository. However, they contain meta-data information, which help their exploitation, for example e-mails.
- Unstructured: data represented or stored without a predefined format.

1.5. Contribution

1.5.1 Meta-model for sewerage networks data sources

Our target is to propose a generic model for data sources, with the aim of using fusion approaches to combine their data. Therefore, on the one hand, our solution should encompass the available approaches for collecting data. On the other hand, the data sources are diverse and may change over time. An exhaustive modelling

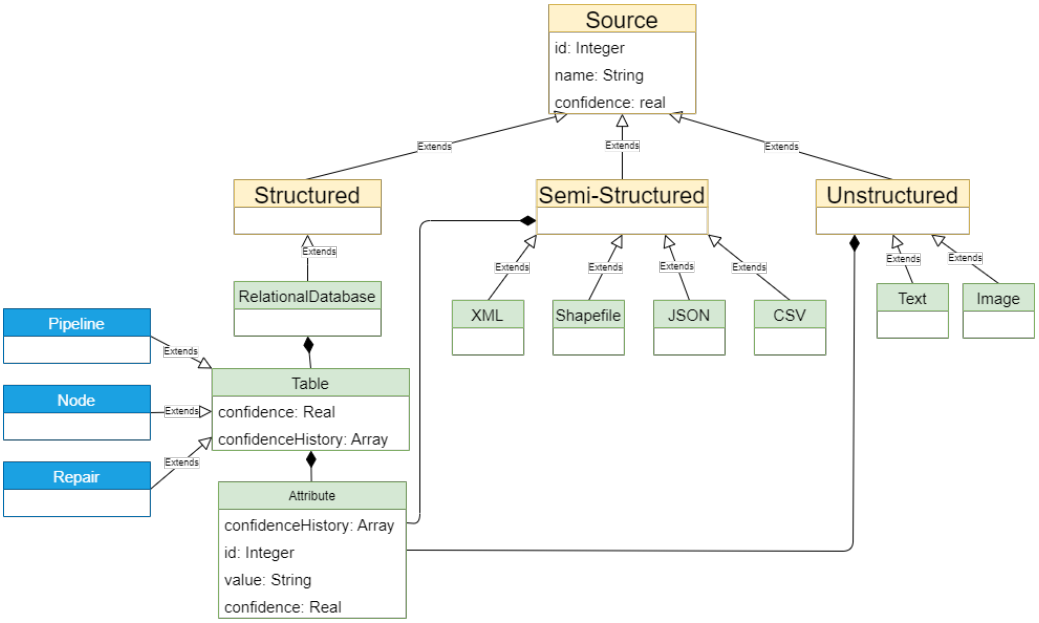


Figure 4 Meta-model for sewerage networks data sources.

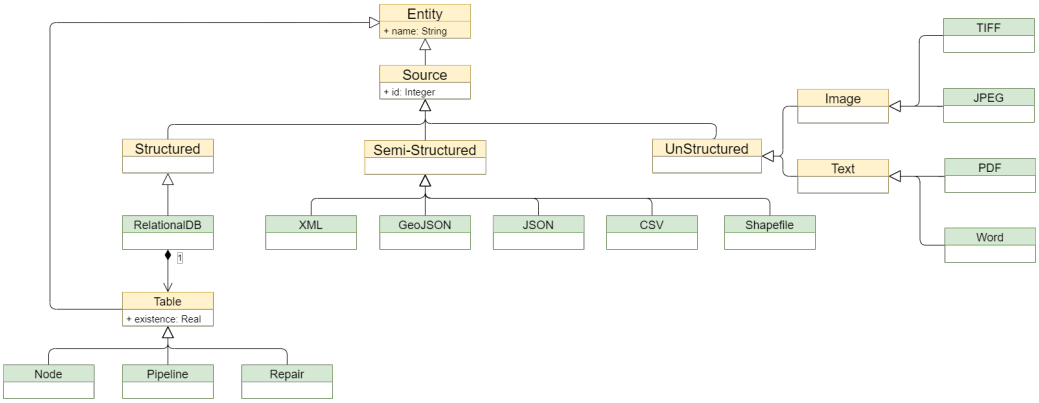


Figure 5 Data sources viewpoint.

of data sources and their possible relationships is not a generic solution. Figure 4 illustrates our meta-model.

For a better understanding, four viewpoints of this meta-model are presented in the following paragraphs.

1.5.2 Data sources viewpoint

We summarise in Figure 5 the data sources viewpoint where the main entity "source" characterises any entity capable of providing data, information or knowledge about sewerage networks. Interpretation of the structuring aspect of data sources is as follows:

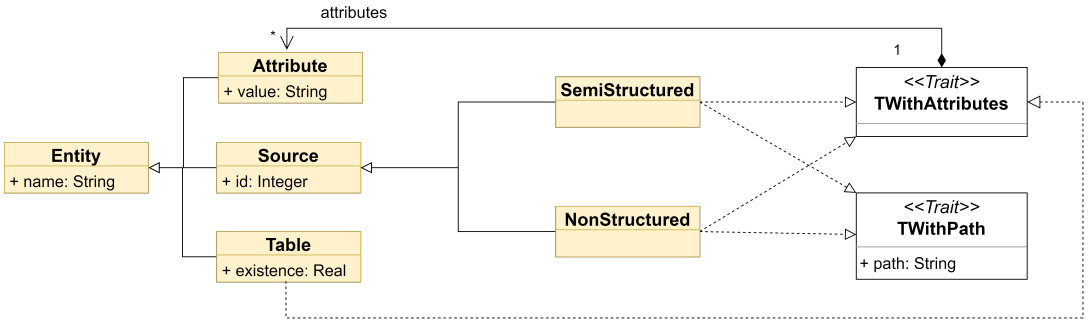


Figure 6 Attributes viewpoint.

- Unstructured sources: whose formats require significant pre-processing before extracting business data about a network. This step generally produces semi-structured data about the networks. For example, a CSV file for the location of sewer manholes detected from images.
- Semi-Structured sources: whose formats require simple pre-processing before extracting business data about a network. For example, parsing CSV or XML files.
- Structured sources: represent relational databases that directly provide business data about a network. Generally, these data are provided by the official managers of sewerage networks.

1.5.3 Attributes viewpoint

The attributes and their relationships with data sources are highlighted in this viewpoint (Figure 6). Each attribute is an entity identified by a name and possesses a String value representing a semantic data. The aggregation of attributes by data sources is encapsulated within the "TWithAttributes" entity. Since it represents, in the case of structured sources, the different attributes within the table of a relational database. Thus, the entity "Table" is connected to this entity. As for semi-structured and unstructured sources, it includes the pre-processing operations, defined by instances of this meta-model, to extract attribute values about sewerage networks. Moreover, data source path is handled within "TWithPath" entity.

1.5.4 Confidences viewpoint

Data imperfection is considered in this viewpoint (Figure 7) by allowing confidence attributes for each source, table (representing an object of the sewerage network) and for each attribute characterizing this table (the object). This means that:

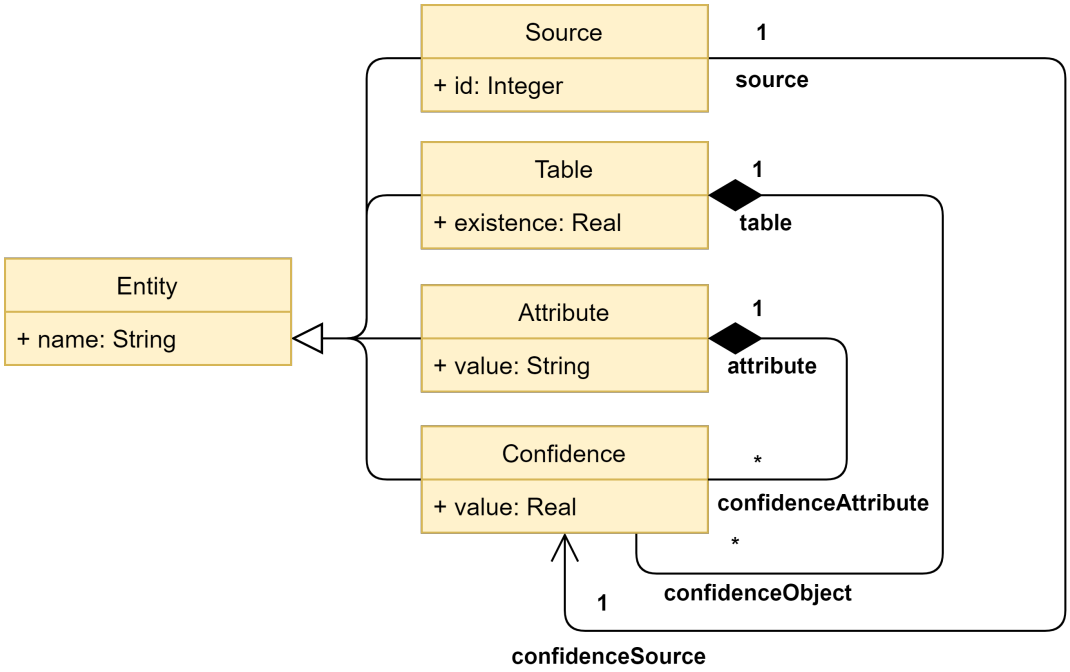


Figure 7 Confidences viewpoint.

- Each source has a confidence value that indicates the reliability or the certainty of the information it provides. This metric can be modelled by the various available mathematical tools, such as probabilities.
- For the components or objects of sewerage networks, this value represents the uncertainty regarding their existence. Indeed, it is possible, for example, for a pipe or a manhole to be represented on a map by mistake.
- As for attributes, the confidence is related to the confidence of the data sources providing them.

Moreover, the objects and attributes default confidence values are those associated with their sources. However, this does not imply not having different confidence values later on. For example, an approach identifying manhole covers from images would define the existence confidence of the detected object as the precision of its detection. Meanwhile, the attribute position of the detected object may have a different value, since the detection and the computation of its position are two separate operations. To keep track of the previous data fusion operations, the confidence history is stored too. This would allow knowledge propagation when new information is collected. Several theories support this propagation such Probability theory and Belief theory (Appriou, 2014).

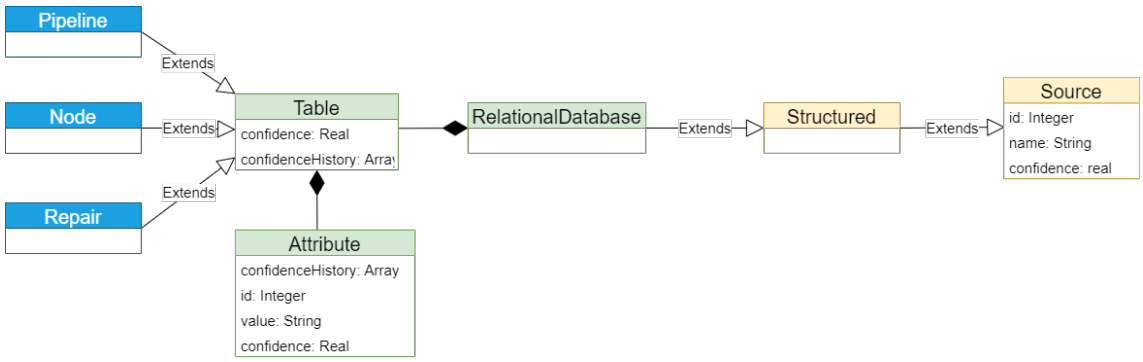


Figure 8 Business model viewpoint.

1.5.5 Business model viewpoint

In our meta-modelling, we distinguish Business data, characterizing the sewerage networks' components, from data about the sources, from which business data are extracted, such as images, documents, calls for tenders, etc.

Figure 8, illustrates this business viewpoint where we adopted the COVADIS (1.4.2) standard classes that inherit the properties of the Table entity. For genericity purposes, other business models may easily replace it, provided that the appropriate connections are respected.

1.6. Use case

1.6.1 Data aggregation for data fusion purposes

Through our proposed meta-model, we aim to perform the fusion of data coming from different sources. In fact, successful fusion operations require a global and complete knowledge about the available data and their sources. To achieve this goal, heterogeneous data, collected from various sources, should be aggregated into a single entity.

According to the data sources viewpoint of our meta-model (Figure 5), we classified the sources as structured, semi-structured and unstructured. Aggregating the attributes' values according to the specificities of each single data source, is considered in the attributes viewpoint of the meta-model within the TWithAttributes and TWithPath traits (Figure 6). Figure 9 depicts the main steps to reach data aggregation from these different sources for fusion purposes.

The first step consists in extracting data from unstructured data sources through automatic and semi-automatic approaches. This usually requires the use of additional

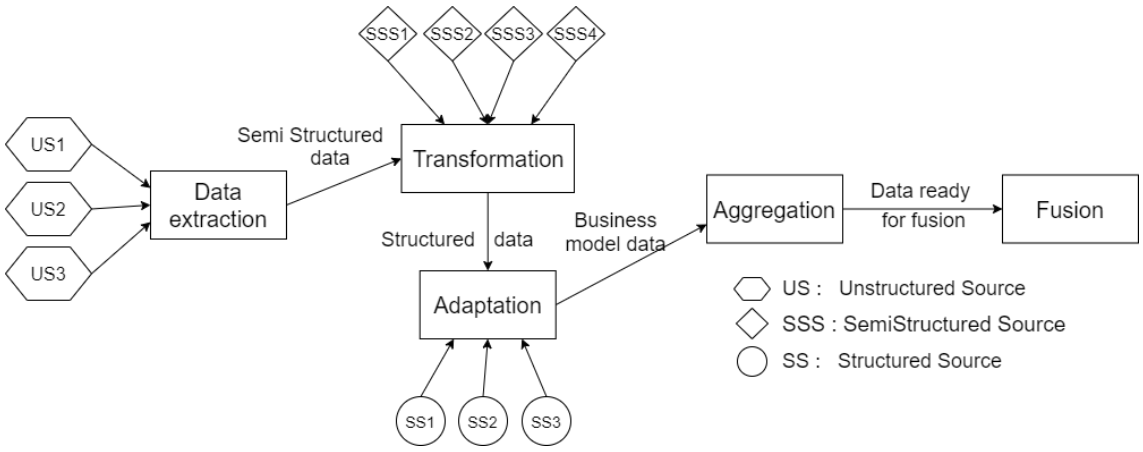


Figure 9 The main steps towards data fusion.

technics related to computer vision, image processing and data mining. Detection of the aerial elements of buried networks from an image is one concrete example. The output of this step is recorded in a semi-structured format, such as flat-file format CSV or JSON.

The second step focuses on the necessary transformations that should be applied to the semi-structured data, together with the available structured data, in order to adapt them to the targeted business data model such as the COVADIS model. These operations of parsing, importing and transforming should be defined for each semi-structured data source and are implemented through what we call a *transformer* for the semi-structured data sources and an *adapter* for the structured data sources. These operations should also include a model manager and an intermediate data visualizer.

Once the data of each single source are pre-processed, the data of all sources could be aggregated allowing to visualize and inquiry them in a unified manner. At this step, the dataset is ready and the fusion operations can be conducted.

To demonstrate this workflow, we conducted an experiment of aggregating data on the sewerage network of Prades-le-Lez, a town located on the outskirts of the city of Montpellier in France, from three different sources. Our first source is the open data database of Montpellier Metropole Médieterrannée, which is a semi-structured source that we note the 3M source. The two other sources are a High-Resolution image database and Google Street View images, which are both unstructured sources that we note respectively HR and GSV sources. To conduct our experiment, we implemented our demonstration using the Moose platform, a software analysis platform.

In the following paragraphs, we will briefly describe the Moose platform and its features. We will then present our different sources and their specific data extraction methods, as well as the data aggregation we performed in the case of a single data source and multiple data sources. Finally, we will report and discuss the obtained results.

1.6.2 Moose

Moose (Ducasse et al., 2005; Moose, 2022) is an open source platform for software and data analysis, currently based on Pharo (Pharo, 2022) a pure object-oriented programming language. It is intended for researchers, engineers, software architects as well as tool builders. The platform is mainly used for software analysis through multiple available mechanisms and features:

- Importing and meta-meta-modelling, since the first step in the process of analysis is the generation of a model of a given target system,
- Parsing, which provides a fluent interface for easy construction,
- Analysis, which provides a rich interface for querying models,
- Visualization through graphs and charts,
- Browsing, which enables the analyst to browse any model.

Moose is also designed to help the programmer build his own tools. This is achieved by the means of several engines through which it can control and customize the complete analysis workflow. In particular:

- build new importers for new datasets,
- define new models to store the data, and
- create new analysis algorithms and tools such as: complex graph visualizations, charts, new queries, or even complete browsers and reporting tools altogether.

Moose was started in the context of the FAMOOS European project (1996-1999), a project focusing on methods and tools to analyze and detect design problems in object-oriented legacy systems, and to migrate these systems towards more flexible architectures. Since then, Moose continued to enhance and integrate new features and tools. Its development is currently supported by multiple research groups, startups and contributors.

1.6.3 Aggregation of a single semi-structured data source

Our first data source is the Open data database of Montpellier Méditerranée Métropole (3M). Through its website (Montpellier Méditeranée Métropole, 2020), this intercommunal structure provides to the public numerous open data files concerning various subjects: transport, finance, environment, etc. The wastewater network datasets are part of these files and include information about nodes and pipes such as the positions of nodes and the dimensions of pipes. We used the dataset related to the position of nodes of the town of Prades-le-Lez town, that we exported as a CSV file.

To get data from a CSV source, we defined an importer MsgMonitoringCSVImporter which takes the path for the CSV file as parameter. The CSV format is popular, thus in Moose platform a default parser for CSV data sources is already defined, the PPCommaSeparatedParser. However, since this parser cannot handle, among other things, white spaces in the values, we have extended it and defined our proper parser: MsgMonitoringCSVParser which takes as input the CSV file content as String value.

Through a stream operation, the importer reads data from the source, handles it to the parser and then passes the resulted parsed data to the CSV model manager: MsgMonitoringCSVModelManager. Since there is no meta-model in Moose for CSV files, the manager designates the parsed output as Intermediate Data Source General Model (IDSGM). Figure 10, illustrates this process.

To visualize the different elements of the MsgMonitoringCSV, we defined the visualizer MsgMonitoringCSVVisualizer whose results are shown in the Figure 11. The left side of the screen shows the used source (file “source1.csv” containing data of Prades-le-Lez) and the right side displays a browser to navigate through the different attributes and their values from the source.

1.6.4 Our unstructured data sources

High resolution images: Extracting data from this source is the result of the work described in (Commandre et al., 2017). The authors used deep convolutional neural network to detect and extract manhole covers from a very high resolution aerial images, which were purchased specifically for their study from a specialized company. This work was conducted on Prades-le-Lez and Gigean, two towns of the 3M (Montpellier Méditerranée Métropole). The Prades-le-Lez dataset was composed of 6 20,000 × 20,000 pixel georeferenced images, and the Gigean dataset was composed of one image of 17,749 × 18,361 pixels. Data augmentation techniques such as rotation, translation and horizontal flip, were used on a small portion of the available datasets in order to train the CNN. The predictions were performed on the remain-

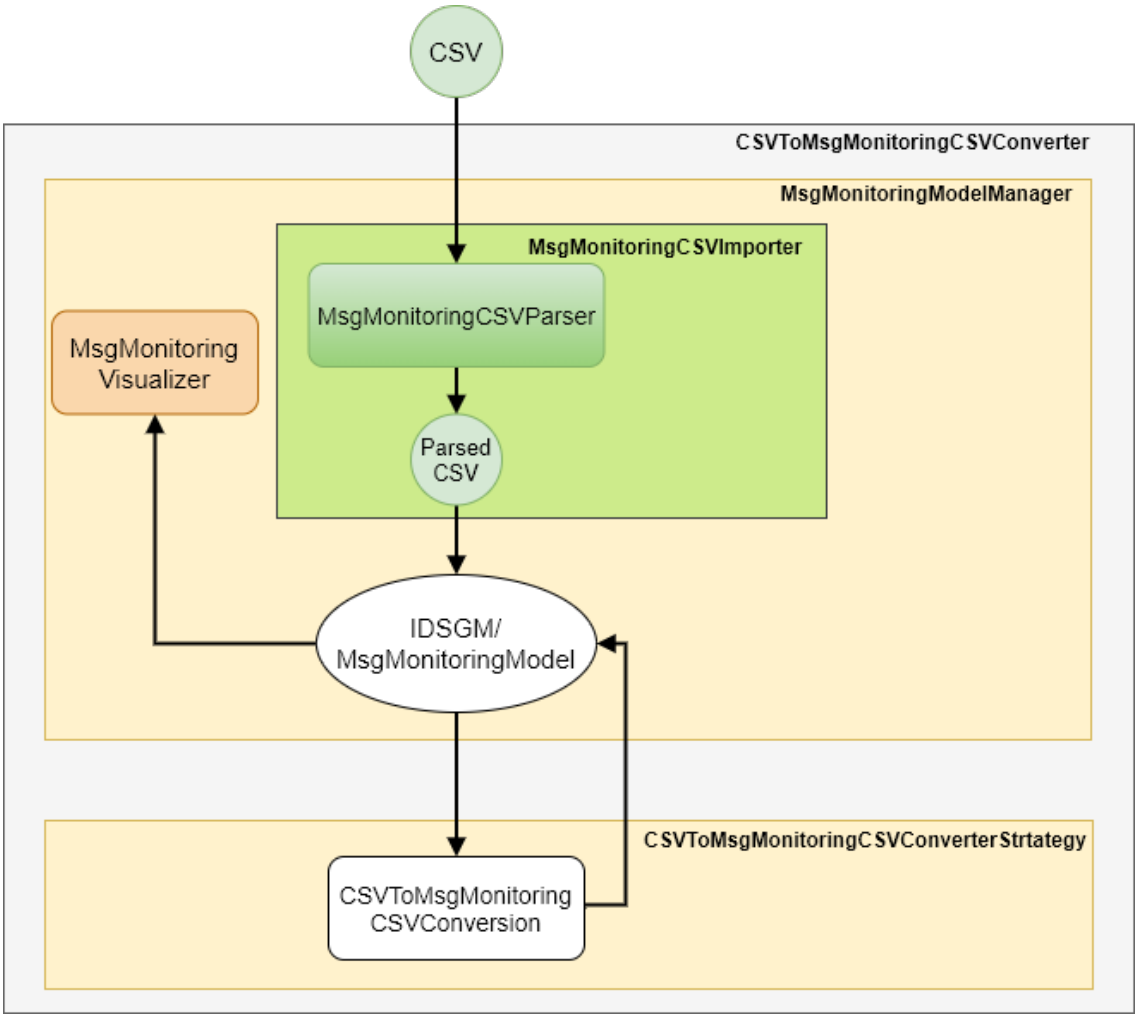


Figure 10 Workflow for the aggregation of a single data source (CSV).

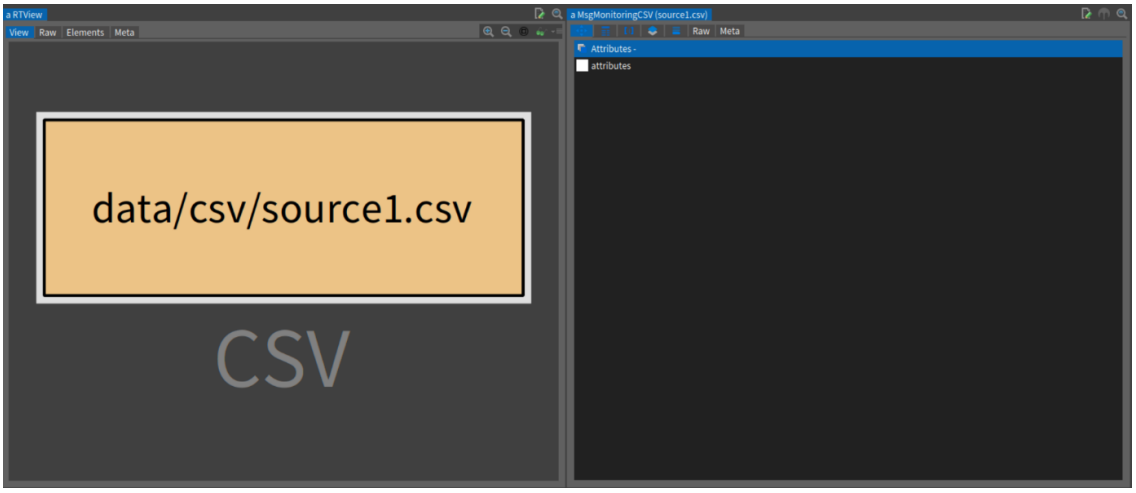


Figure 11 Visualization of the aggregation of a single data source (CSV).

Figure 12 Example of manhole covers' detection using Google Street View images.

ing unlabeled portion. The validation phase was done manually by an expert. The authors were able to detect 49% of ground truth manholes with a precision of 75%.

For the purpose of our experiment, we used the CSV semi-structured format to store the output of the manhole detection from high resolution images, and thus we used the workflow described in Figure 10 to aggregate the data from this source in Moose.

Google Street View Images: Considering that in reality, a large part of the manhole covers are located on roads and streets, we propose as a second unstructured source for our experiment, to detect manhole covers from Google Street View Images. Indeed, they provide an interactive panorama of the majority of streets in France, and many countries around the word. To get the positions of manhole covers of Prades-le-Lez, we proceeded by collecting all the images of Prades-le-Lez corresponding to the nodes positions provided by our first, official data source (3M). We proceeded this way in order to reduce the number of unnecessary images to be treated.

The following 3 main steps summarize our data collection process:

- Using Google Static Street View API, we were able to collect images corresponding to the positions provided by 3M.
- To detect manhole covers, we used the You Only Look Once (Yolo) object detection algorithm (Redmon & Farhadi, 2018), that we trained on thousands of manhole covers.
- We exported the positions of the detected manhole covers along with the precision of the detection to an XML file. Figure 12, illustrates an example of a detected manhole cover from a Google Street View image.

We used XML as a semi-structured storage format for this data source to demonstrate the usefulness of the proposed aggregation of data from multiple semi-structured formats. Therefore, we defined an XML workflow (Figure 13) similar to the CSV workflow (Figure 10). In other words, we extended the default Moose Parser for XML PPXmlParser to MsgMonitoringXMLParser, and defined the different components necessary for the workflow: the importer MsgMonitoringXMLImporter, the model manager MsgMonitoringXMLModelManager, the conversion XMLToMsgMonitoringXMLConversion, the strategy XMLToMsgMonitoringXMLConversionStrategy and the visualizer MsgMonitoringXMLVisualizer. The visualization for the different sources is managed by the MsgMonitoringVisualizer.

1.6.5 Aggregation of multiple data sources.

To aggregate data from multiple sources in Moose, a coordinator is necessary. Therefore, we defined ModularIDGSMToMsgMonitoringConverter, which contains the different convertors prescribed for each data source as modules. The role of this coordinator is to choose automatically for each data source, the corresponding converter. The coordinator receives data sequentially from different data sources, then for each iteration, the data source associated module is identified and used to trigger the workflow explained bellow. That is to say, generating the IDSGM model from the parsed data using the importer, and then aggregating data to the associated model from the MsgMonitoringModel (Figure 13).

Figure 14, shows the visualization of our 3 semi-structured data sources in Moose, two CSVs (the 3M and HR sources) and one XML (the GSV source), managed by the MsgMonitoringVisualizer.

To finalize our demonstration, we developed an API for the analysis, interaction and monitoring of the aggregated data provided by the different sources, through a series of queries and requests. To achieve this goal, we based our API on the design patterns Visitor and Strategy. Our meta-model has a hierarchical structure starting from the global entity source to the single attribute values. Consequently, we defined three levels of the analysis:

- Source level analysis: where the available sources and their specific information like their providers and the associated confidence values are listed. We used the Visitor design pattern to implement this level.
- Object level analysis: using also the Visitor design pattern, this level lists the different available objects for each source. For example, the 3M source

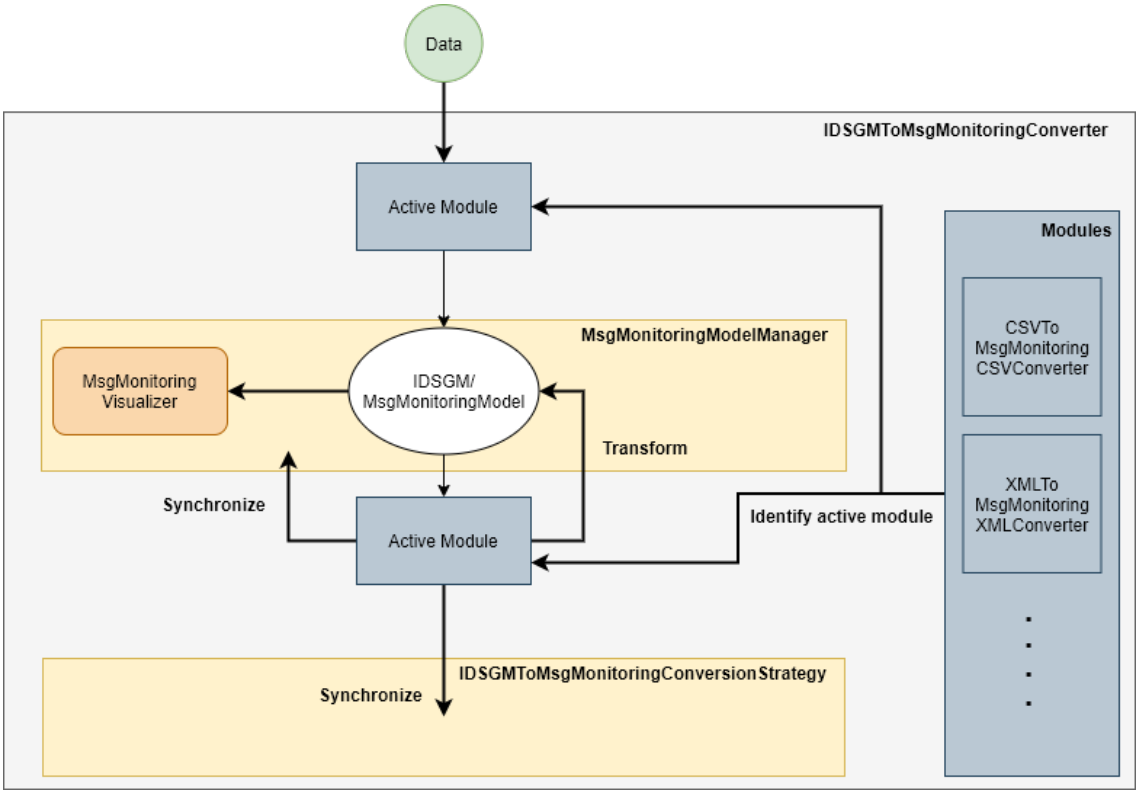


Figure 13 Workflow for the aggregation of multiple data sources (CSV and XML).

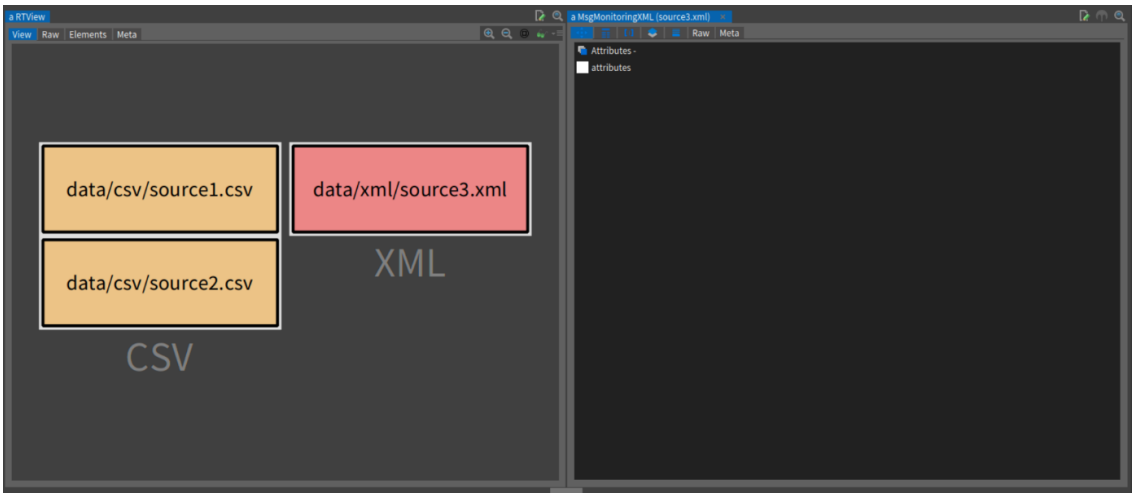


Figure 14 Visualization of the aggregation of three data sources.

provides data about nodes and pipes, whereas the GSV source produces data about nodes only.

– Attribute level analysis: this level is more sophisticated, since it processes data sources in their deep details, the attributes and their values. Thus, besides the Visitor design pattern we used Strategy which allows defining for each data source, the specific required operations for their analysis and monitoring. For example, for a CSV data source, and in addition to the elementary requests of attributes values, the user may request the type of separator, the available attributes, the distribution of attributes values, the line at a specific index, the first n lines, lines between two indexes, etc.

Table 1 Examples of queries using the three sources.

(a) Confidence values of the sources.

Sources	CSV_3M_source	XML_GSV_source	CSV_HR_source
Confidence	0.90	0.60	0.75

(b) Confidence values of the attribute position.

CSV_3M_source			XML_GSV_source			CSV_HR_source		
ID	Position	Confidence	ID	Position	Confidence	ID	Position	Confidence
613	3.8632613,43.6916825	0.90	78	3.8632613,43.6916825	0.75	12	3.8632613,43.6916825	0.76
622	3.8642639,43.7010732	0.90	80	3.8642639,43.7010732	0.58	52	3.8642639,43.7010732	0.80
623	3.8660806,43.6890539	0.90	45	3.8660806,43.6890539	0.98	32	3.8660806,43.6890539	0.84
602	3.8653489,43.6939405	0.90	14	3.8653489,43.6939405	0.96			
617	3.8629708,43.6884132	0.90	19	3.8629708,43.6884132	0.94			
						66	3.8734563,43.6932746	0.82
						207	3.8646406,43.7005578	0.80

1.6.6 Results and discussion

Table 1a, illustrates a concrete example of the source level analysis where we list the three sources used in this experiment and their confidence values. This may allow the final user to make conclusions for decision-making or conduct some reasoning to understand the influence of these values on the confidence of other inter-related data, or infer them in deeper levels of analysis. As we mentioned before, the perspective of our work is to automate this reasoning task for fusion purposes. In the current experiment, we can explain the confidence values we assigned to the sources as follows:

– The 3M source is a priori reliable, as it was provided by the official managers of the network. However, on the one hand, data are incomplete as not all of the attributes and their values are listed in the attribute tables. On the other hand, considering that sewerage networks evolve in time, there is no guarantee that this source has continuously updated its data during its years of service. Therefore, taking into consideration all of these facts, a

subjective confidence value of 0.9 was assigned to this source by a domain expert.

- We assigned to the HR source the value of 0.75 as an objective confidence. It corresponds to the "precision" metric of the algorithm of the detection of manholes covers, namely 75%.
- The Google Street View Images: when we collected Google Street View images we used only the node positions provided by 3M. However, not all those positions were associated to an image. In fact, among the 799 positions that we used, we found only 763 corresponding images, more than half of them didn't contain a manhole cover, which is due to the presence of diverse objects on the street such as cars hiding manholes, machine learning model failing to identify manholes covers, also considering that not all the node positions provided by 3M represent manholes covers but they can represent junctions of underground pipes. Therefore, based on this information we assigned objectively the value 0.6 (451 / 763) as the confidence of this source.

At the object level analysis, we could display the 1,862 collected nodes from the 3 sources: 799 nodes from the 3M source, 451 nodes from the GSV source and 612 from the HD images source. Even if we used the node positions provided by 3M to collect data from the GSV source, only 451 of the 799 nodes have been detected.
We aggregated data from the 3M, the GSV and finally the HR source respectively. We categorized the nodes at the attribute level analysis, as appearing in the 3 sources, in 2 from the 3 sources or only in one source. This information is important in the process of data combination to confirm the positions of the nodes previously available in our records and to detect potential nodes that were not identified.
Table 1b. shows examples of nodes where the attributes and their confidence values are listed. It is to be noted that the confidence values we assigned to the attributes represent the precision value of the manhole cover detection from the images; which explains the differences between the confidence values of the nodes in the same source for the HR and the GSV sources. Since the 3M sources is a semi-structured one, there had been no need to the extraction step and then all the nodes inherited the confidence value of their source.

The first 3 lines of the table represent 3 nodes appearing in the 3 sources (corresponding to 9 detected nodes from the 3 sources). The 2 following lines are nodes appearing in both the 3M and the GSV sources, which allows to confirm their positions. The last 2 lines represent nodes detected only by the HR source. In this case, they may represent an extension of the network that is still not reported by the official source,

thus really missing data, or they may be simply false positive cases extracted using the HR source.

Although, the results of the detection from the high resolution images and Google Street View images are far from perfect, this example demonstrates that they can be used to confirm and update the confidence value of the attributes of the 3M source. Such images can also be used to create a fourth dataset where the objects appearing in only one source (Section 1.6.6) would have low confidence values and those in common would have a fused confidence value.

1.7. Conclusion

In this chapter, we proposed a meta-model for data sources related to sewerage networks inspired from the field of Big Data. The main objective is to perform a multi-source data fusion on sewerage networks to make a more complete dataset available to the decision makers, taking into account data imperfections. Our proposition considered the three important aspects of *i*) structure of the data sources: structured, semi-structured and unstructured, *ii*) associated confidences at multiple levels and *iii*) genericity related to the business domain. As a first step towards sewerage networks data fusion and through our meta-model, we implemented this very initial step in Moose, a platform for software and data analysis. As a concrete example, and to show that our meta-model is generic and able to encompass the various sources and approaches for the collection of data about sewerage networks, we used one semi-structured source and two unstructured sources with their appropriate data extraction processes. The results enabled the listing, the visualization and the analysis of the aggregated data which suffer in most cases from imperfections. We are currently conducting some experiments on spatial sewerage data fusion that we expect to enhance by proposing a data fusion approach with uncertainty management allowing to cop with the multi-source data imperfections. A validation example was provided in the previous section. It should however be noted that additional case studies are necessary to insure the genericity of our proposition.

2. Object Matching based on Dempster-Shafer's Theory

In the previous chapter, we set up the foundation for the data fusion of wastewater network data sources. We proposed a generic meta-model for wastewater data sources and we showed an example of data visualisation and analysis of the sources. In this chapter, the heart of the data fusion process is described, where the actual combination of the data sources is accomplished. We focus on object matching, precisely on identifying corresponding pipes from multiple data sources. We propose a line matching process, driven by the stroke concept in order to capture the structure of the networks and thus, reduce the impact of missing data on the matching results. To take into consideration the source imperfections, the similarity measures are combined using Dempster-Shafer's theory. This chapter is an answer to the second and the fourth scientific questions formulated in the general introduction. This study is submitted and under review in the International Journal of Approximate Reasoning.

2.1. Introduction

Spatial data collected from different sources, even though they may be incomplete and uncertain, can be used in a combination process to improve the quality of the dataset by: *i)* confirming similar data, *ii)* completing missing data and *iii)* detecting inconsistencies. 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 (Devogele et al., 1998). Data conflation is defined as the process of creating a new dataset based on multiple datasets that cover the same spatial area (Tong et al., 2009). 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 and conflation (Costes & Perret, 2019; Song et al., 2011; Volz, 2006).

Object matching is the task of identifying homologous objects, that represent the same real entity, from two or more spatial databases (Tong et al., 2009; Volz, 2006). The objects to be matched are usually represented by vector structures and classified in three categories: nodes, lines and polygons. Historically, the US Census Bureau was one of the first to initiate matching objects in order to achieve map conflation of separate digital maps of metropolitan areas (Saalfeld, 1988). Since then, several studies have been published and significant progress has been achieved.

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 also especially because of data imperfections such as incompleteness, distortion and imprecision. Managing the imperfection in object matching comes down to answering two questions: *i*) How to deal or model the imperfections? *ii*) How to reflect these imperfections on the matching results? To deal with imperfections, objects from different datasets are compared using different criteria such as the position or the form, and 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 suffers from important distortion, the contribution of a distance based criterion is less important than the one based on the topology. Traditionally, the contribution of each criterion is modeled by weight values, and the uncertainty or the confidence of the output matching is computed as a weighted mean of all the chosen criteria. Besides this basic approach, some studies like (Tong et al., 2009), proposed to use the probability theory to produce more representative uncertainty of the matching results. The probability theory is designed to deal only with uncer-

tainty and cannot explicitly handle other imperfections such as data imprecision and sources reliability. Dempster-Shafer's (DS) theory (Dempster, 1967; Shafer, 1976), one of the most advanced approaches to deal with data imperfection (Appriou, 2014) was applied in (Deng et al., 2019; Nassreddine et al., 2009; Raimond et al., 2015) to achieve object matching for roads and Point of Interest applications.

In this chapter, we aim to achieve object matching of wastewater spatial data, taking into account data imperfections. Our matching proposition is guided by the DS theory, which formally models uncertainty, imprecision and reasoning under incompleteness constraints. This contribution is structured as follows: section 2.2 gives an overview of the proposed methods in the literature. Section 2.3 presents background on DS theory. Our proposition for matching objects of wastewater networks is introduced in section 2.4. The tests and the results are presented and discussed in sections 2.5 and 2.6. We conclude this work in section 2.7.

2.2. Research background

Object matching is conducted following different approaches and methods which generally share two main steps: *i)* identifying similarity measures and *ii)* defining a matching approach that use these measures. In the following, a set of similarity measures and a variety of matching approaches are described.

2.2.1 Similarity measures

To assess whether two or more spatial objects from different databases represent the same real entity, comparison criteria – referred to as similarity measures – are required and form an essential part of any approach. The most common criterion is the one based on position, where two objects are supposed to be homologous when they are close in terms of a given distance. The most popular one is the euclidean distance (Beeri et al., 2004; Samal et al., 2004; Volz, 2006). However, when spatial objects of two or more databases are characterised by important deviations and distortions, one may use other criteria such as geographic context or attributes. Similarity measures and distances are complementary concepts. Generally a distance is converted to a similarity by normalizing its value (Samal et al., 2004). Hence, two objects are more likely to correspond when the similarity measure is high. In the literature, similarity measures are usually divided into three categories:

Geometric: measures 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 (Beeri et al., 2004).
- 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 (L. Li & Goodchild, 2011; Volz, 2006; Xavier et al., 2016). Given two points 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 Hausdorff distance, the Fréchet distance considers the ordering of the points along the shapes which make it suitable for comparing curves, but harder to compute (Alt et al., 2004). A small outlying point from one points set lead to an important increase in the Hausdorff distance. To minimize this impact of outlying geometry, an extension of the Hausdorff distance based on the median Hausdorff distance was proposed in (Min et al., 2007).
- To compare the form of the geometry, the length is used for lines as in (Walter & Fritsch, 1999), and the area for polygons as in (Samal et al., 2004).
- The angle is usually used to measure the orientation of lines as in (Raimond et al., 2015). It is computed by the differences in angles or directions of two or more objects compared to a fixed axis (Samal et al., 2004; Tong et al., 2009; Walter & Fritsch, 1999).

In addition to these popular geometric distances, others can be developed and used such as the perimeter of a triangulation, as proposed in (Kim et al., 2010).

Topological: captures the relationships between an object and its neighbours. The node degree measure, i.e. the number of edges connected to a node, is one 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 roads matching (Saalfeld, 1988; Song et al., 2011; Volz, 2006). Several other topology-based measures were proposed. For example, in (Samal et al., 2004), 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 (S. Wang et al., 2019), a spatial scene composed of the neighbours of the objects is used as topological measure to match underground networks.

Attribute metrics: compares 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 (Bilenko & Mooney, 2003; Costes, 2014; Samal et al., 2004), defined as the minimum number of insertions, deletions and substitutions needed to transform one character string to the other. In addition to the value of the attributes, their semantic is often used when it's necessary. For example, in (Raimond et al., 2015) a conceptual similarity between object types is analyzed to minimize aberrant matching such as matching a valley with a summit. In (Assi & Dhifli, 2021), a semantic-similarity step, based on word embedding, was used to achieve instance matching, and was applied on real world datasets such as restaurants, health and drug domains.

Distances are generally turned into similarity measures without taking into consideration candidates ranking which means that when the distances are almost equal, so are the similarity measures. However, since we seek to match objects that are the most similar, the closest candidate should be emphasized over the second closest one (even if the difference is small) when defining the similarity measures. In (Beeri et al., 2004), to account for candidates ranking in nodes matching, a probability value is derived from the Euclidean distance as follow:

$$\mathrm{P}_{\mathrm{a_i}}(b_j) = \frac{d(a_i, b_j)^{-\beta}}{\sum_{k=1}^{\mathrm{N_c}} d(a_i, b_k)^{-\beta}} \tag{2.1}$$

where N_c is the number of candidates b_k to be compared with node a_i, d is the Euclidian distance, β is a parameter to decrease the measure when the distance to a_i increases. The more β increases the more the gap between the closest and the other candidates increases. Hence, the ranking of the candidates based on the Euclidean distance is reflected on the probability value.

2.2.2 Object matching approaches

Object matching has been performed in numerous domains such as road networks (L. Li & Goodchild, 2011; Tong et al., 2014; Volz, 2006), hydrographic networks (Costes, 2014), underground networks (S. Wang et al., 2019) or buildings (Beeri et al., 2004; Y. Wang et al., 2015). The used approaches can be classified according to different criteria. In the following paragraphs, we give an overview of these studies according to the supported constraints and to the matching steps.

2.2.2.1 Classification according to the supported constraints

To choose the suitable methods among the available propositions, a user must verify if a method satisfies the constraints of the datasets at hand. Each proposed approach in the literature can be characterised by the four following supported constraints:

- The supported objects' representation: nodes (e.g. (Beeri et al., 2004)), lines (e.g. (S. Wang et al., 2019)) or polygons (e.g. (Kim et al., 2010)). Some propositions support all three representations as in (Tong et al., 2009).
- The supported representation scale: some studies address datasets with identical scales (Beeri et al., 2004; Walter & Fritsch, 1999) and others with different scales (Devogele et al., 1998; Kim et al., 2010; Y. Wang et al., 2015).
- The supported cardinalities: in general we distinguish between the methods that accept only (1:1), (0,1) 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 (Kim et al., 2010; Wu et al., 2018).
- The number of datasets that can be matched: methods can support either 2 (e.g. (Kim et al., 2010)) or N> 2 (e.g. (Samal et al., 2004)) datasets.

2.2.2.2 Classification according to the matching steps

As indicated before, all matching methods use similarity measures to discriminate between the possible candidates to match. Except for cases like in (Beeri et al., 2004), where the distance between nodes is the only available measure, most matching methods use various measures. Therefore, what differentiates them is how they define, use and combine these measures. First, we distinguish between general methods which are independent from the measures (i.e. the distances are just parameters to tune), and those in which the measures are part of the matching process as in (Costes, 2014), where the authors use Horton classification and pre-matched nodes to match hydrographic networks. A second difference is that the matching can be either mono-directional, from one dataset (generally called reference), to another dataset (called target) (Kim et al., 2010), or bidirectional where the datasets are used simultaneously as reference and target (Y. Wang et al., 2015). The matching process can be summarised in 3 main steps:

- Candidates' selection: consists in selecting the closest objects, since corresponding ones are usually spatially close. 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 (Costes, 2014; Kim et al., 2010; L. Li

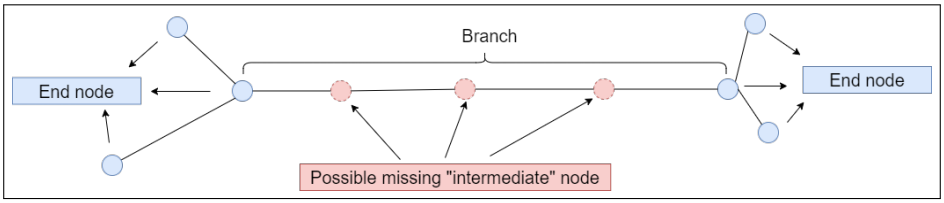


Figure 15 Example of missing nodes of a wastewater network, where the end nodes are always available and the nodes within the branches ("intermediate") are the ones that may be missing.

& Goodchild, 2011; Samal et al., 2004; Volz, 2006). As for filters, the candidates are generally those having a measure value lower than a fixed threshold (Beeri et al., 2004; Costes & Perret, 2019; L. Li & Goodchild, 2011; Samal et al., 2004; Y. Wang et al., 2015; Wu et al., 2018).

- 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 (Samal et al., 2004; Tong et al., 2009; Y. Wang et al., 2015; Zhang et al., 2005). 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 (Y. Wang et al., 2015), where the weights are learned using an Artificial Neural Network (ANN). Another combination scheme consists in decreasing the filters' thresholds iteratively until only one candidate at most remains (Zhang et al., 2005).

- Decision: usually the candidate with the highest score is chosen as a match, however optimisation based approaches are also popular: for example in (Samal et al., 2004), the authors transformed the decision step into the maximal clique problem. In (Walter & Fritsch, 1999), the correspondence is encoded onto a communication system where the transmitter is a reference dataset, the receiver is the target dataset and the channel which transmits the message from reference to target is the mapping function.

2.2.3 Towards matching wastewater network

In the graph provided by some data sources, data incompleteness may occur, like missing nodes. Indeed, wastewater networks are composed of a set of connected pipes. Sometimes, multiple contiguous pipes can be represented by only one line. Thus, intermediate nodes indicating the location of the junction between the pipes

are not recorded (Figure 15). This can be due to several reasons: *i*) lost data *ii*) the nodes are irrelevant for certain applications as long as the overall structure of the network is preserved by few nodes, such as the case for hydraulic simulations *iii*) the limitation of the method used to collect the data, such as radar based approaches. In addition, road, river or 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 be exploited to handle the issue of missing nodes. To consider this aspect, in (Costes, 2014; S. Wang et al., 2019; Zhang et al., 2005), the authors propose to use the concept of stroke as a matching unit, where a stroke describes a "good continuity" between lines that do not exceed a degree of deviation. 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.
Missing nodes represents only one aspect of data imperfections. In the litterature ((Beeri et al., 2004; Bordogna et al., 2010; Raimond et al., 2015; Rosen & Saalfeld, 1985)), data imperfections that could be encountered in the matching process can be related to the: *i*) reliability of the sources; *ii*) reliability of the similarity measures; *iii*) uncertainty and incompleteness of the attribute values, where uncertainty is related to the veracity of an assertion and incompleteness designates the absence of data; iv) imprecision and uncertainty of the produced matching, where imprecision is the doubt about multiple possible values. Usually, at most only two of these four aspects are considered, and often using basic operations which do not model all aspects of the imperfection. In (Beeri et al., 2004), the confidence of the matching of two object a and b was computed as follows:

$$\text{confidence}(a,b) = 1 - \frac{\text{distance}(a,b)}{\min\{\text{distance}(a,b'), \text{distance}(a',b)\}} \tag{2.2}$$

where a' is the second-nearest neighbour of b after a, and b' is the second-nearest neighbour of a after b. In (Samal et al., 2004), an optimization method was applied to identify the matching objects, and the system's confidence in the output is computed. This value was measured as the sum of the total number of features in each similarity set divided by the total number of features being matched. The reliability of the measures is often considered by assigning a weigh value to each measure, such as in (Y. Wang et al., 2015) and (Tong et al., 2009).

Compared to other mathematical theories such as probability, DS theory provides suitable tools to handle combination of similarity measures under imperfections constraints and has been used in (Deng et al., 2019; Nassreddine et al., 2009; Raimond et al., 2015) for this purpose. Although, available approaches such as optimisation methods can produce efficient matching results, they were not designed to handle

data imperfection properly. In this work we use DS theory to match imperfect data of wastewater networks by combining several similarity measures computed from distances between the objects of different sources, and by modeling data uncertainty and imprecision under the constraints of incompleteness and reliability of the sources. In the next section, we present the main concepts of this theory and review the related studies for object matching.

2.3. Dempster-Shafer's theory

2.3.1 Main concepts of Dempster-Shafer theory

We consider the frame of discernment $\Omega = \{H_1, H_2 \dots H_N\}$, that denotes the set of the discrete, mutually exclusive and exhaustive possible N events (Hypothesis). The likelihood associated to a subset of Ω is defined by the mass function $m(.)$:

$$m : 2^{\Omega} \mapsto [0, 1] \text{ with } \sum_{A \subseteq \Omega} m(A) = 1 \tag{2.3}$$

where A is called the focal element when $m(A) > 0$. The values of $m(.)$ model the uncertainty. 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 $A = \{H_1, H_2, H_3\}$ is not equally distributed between the singletons H_1, H_2 and H_3, contrary to the probability theory. Hence, the value associated to m(Ω) represents the ignorance and not an equiprobability.

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

$$\text{Cr}(\varnothing) = 0, \ \text{Cr}(\Omega) = 1 \text{ and } \text{Cr}(A) = \sum_{B \subseteq A} m(B) \tag{2.4}$$

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

$$\text{Pl}(A) = \sum_{A \cap B \neq \varnothing} m(B) \tag{2.5}$$

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:

$$(m_1 \cap m_2)(A) = \sum_{B \cap C = A} m_1(B) \times m_2(C), \quad A \subseteq \Omega \tag{2.6}$$

– Normalized conjunctive rule of combination:

$$m_1 \oplus m_2 = \begin{cases} \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) \times m_2(C) & \text{if } A \neq \varnothing \\ 0 & \text{if } A = \varnothing \end{cases} \tag{2.7}$$

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

$$K = m(\varnothing) = \sum_{B \cap C = \varnothing} m_1(B) m_2(C) \tag{2.8}$$

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

$$m_1 \cup m_2(A) = \sum_{B \cup C = A} m_1(B) \times m_2(C), \quad A \subseteq \Omega \tag{2.9}$$

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 $R(H_i)$ 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:

$$\forall A \subseteq \Omega^1 \; m^2(R(A)) = m^1(A)$$

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:

$$m^\alpha(A) = (1 - \alpha) m(A) \tag{2.10a}$$

$$m^\alpha(\Omega) = (1 - \alpha) m(\Omega) + \alpha \tag{2.10b}$$

Belief theory offers several rules of decision, such as selecting the hypothesis with the highest belief, plausibility or pignistic probability value. The pignistic probability is defined in (Smets & Kennes, 1994) as follow:

$$\mathrm{BetP}(H_i) = \sum_{A \subseteq \Omega, H_i \in A} \frac{m(A)}{|A|} \tag{2.11}$$

where $|A|$ is the cardinal of A.

2.3.2 Dempster-Shafer theory in object matching

In the context of object matching, DS theory has been previously applied in (Deng et al., 2019; Nassreddine et al., 2009; Raimond et al., 2015). To estimate accurately the position of a moving vehicle on a road, the authors in (Nassreddine et al., 2009) used the normalized conjunctive rule to combine three different mass functions. Only one of these three functions was computed based on a similarity criterion, following three steps. First step: a likelihood value $\mathrm{L_i}, i \in \{1 \dots N\}$ was determined based on the intersection between the current vehicle position and the N candidate roads. Second step: the mass values of the frame of discernment $\Omega_i = \{\mathrm{H_i}, \overline{\mathrm{H_i}}\}$ were initialized based on the following model:

$$m_i = \begin{cases} m_i(\{\mathrm{H_i}\}) = 0 \\ m_i(\overline{\{\mathrm{H_i}\}}) = r_i(1 - \mathrm{L_i}) \\ m_i(\Omega) = 1 - r_i(1 - \mathrm{L_i}) \end{cases} \tag{2.12}$$

where H_i is the hypothesis of being the road i, r_i is a discounting coefficient associated to H_i, and $\overline{\{\mathrm{H_i}\}}$ is the complement of $\{H_i\}$ in Ω. Last step: the mass functions of the candidates in Ω, were combined using also the normalized combination rule. The decision was taken based on the highest pignistic probability.

Similarly to the previous work, the authors in (Deng et al., 2019) combined multiple criteria using DS theory to match spatial objects (Points of Interest). To reduce the conflict C generated by the conjunctive rule, a modified likelihood initialization is proposed. Contrary to the previous proposition, where the mass functions initialization related to the frame of discernment $\Omega_i = \{\mathrm{H_i}, \overline{\mathrm{H_i}}\}$ set the value of $\mathrm{H_i}$ to 0, in this study the model applied to transform the likelihood into mass values is the following:

$$m_i = \begin{cases} m_i(\{\mathrm{H_i}\}) = r_i(\mathrm{L_i}) \\ m_i(\overline{\{\mathrm{H_i}\}}) = r_i(1 - \mathrm{L_i}) \\ m_i(\Omega) = 1 - r_i \end{cases} \tag{2.13}$$

where $r_i = 1$.

In order to find corresponding objects, a multi-criteria data matching approach guided by a formal representation and fusion of sources using the DS theory is proposed in (Raimond et al., 2015). After extracting candidates for each reference object, a set of distances are computed and transformed into mass functions using the following model:

$$m_i = \begin{cases} m_i(\{\mathrm{H_i}\}) = v_1 \\ m_i(\overline{\{\mathrm{H_i}\}}) = v_2 \\ m_i(\Omega) = v_3 \end{cases} \tag{2.14}$$

where v_1, v_2 and v_3 are initialized based on a subjective knowledge about the supported 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 (Appriou, 1991), a two level fusion step is carried out to decide which couples match: first, for each potential couple, the initialized mass functions based on the similarity distances are combined using the normalized combination rule, second, based on the refinement operation, the mass functions of all the potential couples resulting from the first step are combined. The decision is made based on the highest pignistic probability.

Although each of these studies has its specificity, the overall steps are similar. We summarize them in (Figure 16). After setting the distances between the reference object and the candidates, the goal of the first step, that we named *transformation*, is to transform these distances into similarity measures or likelihood values, that are used to initialize the mass functions.

Regarding this transformation, A. Appriou in (Appriou, 1998, 2014) distinguished between two models in which the above studies fall. The first one (equation 2.12), is based on the assumption that when the likelihood $\mathrm{L_i}$ of a criterion (in our case a similarity measure) about a hypothesis $\mathrm{H_i}$ (in our case a candidate) is equal to 0, we are sure that the hypothesis is not verified ($m_i(\overline{\{\mathrm{H_i}\}}) = 1$ when $r_i = 1$). However, the same criterion could serve to identify multiple hypotheses, and is not specific to only one hypothesis, thus the value of the likelihood $\mathrm{L_i} > 0$ is assigned to the ignorance mass value $m(\Omega)$.

When a criterion represents a signature for a hypothesis, one may use the second model (equations 2.13 and 2.14) where the likelihood is directly assigned to the mass value of the hypothesis. Therefore, the choice of the transformation model must be justified and adapted based on the available data.

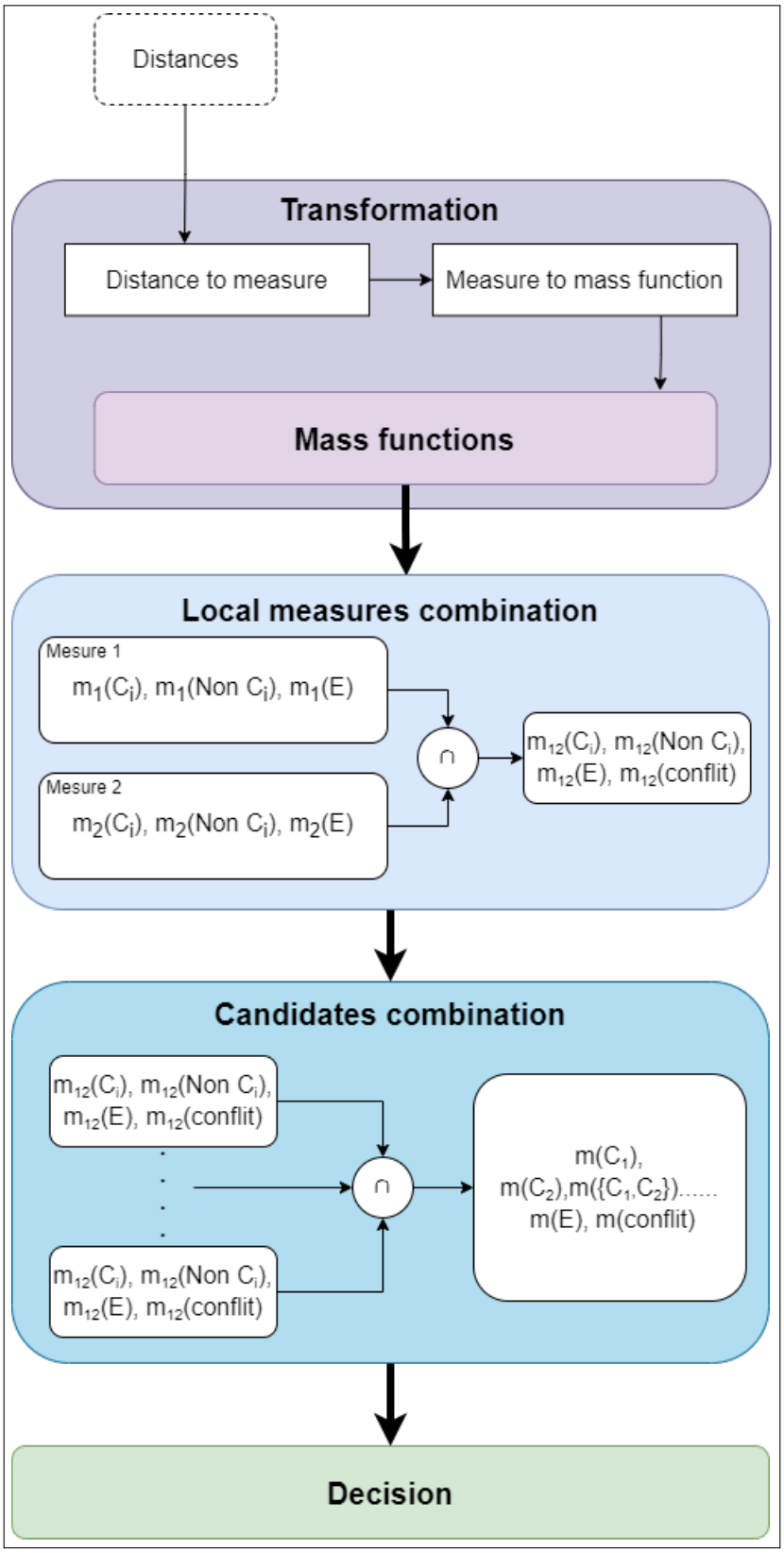


Figure 16 Object matching process using the DS theory.

The second step is to combine the mass functions locally for each couple formed by a reference object and a candidate (see *local measures combination* step in Figure16). Next the resulted mass values are combined between the candidates (see *candidates combination* step in Figure16). Finally a decision is taken based on various characteristics such as plausibility or the pignistic probability.

2.4. Materials and methods

In this work, we aim to match the spatial objects of wastewater networks collected from different sources: wastewater pipes represented by lines. This includes dealing with uncertainty, imprecision and incompleteness. We describe our matching approach using the stroke concept and the DS theory. We also show that our method can be generalized to other applications. This section is organized as follows: first we introduce the four distances metrics used to evaluate the similarity between the lines, second we describe our matching process, then we detail our proposition for multi-criteria decision making based on DS theory.

2.4.1 Similarity distances

In order to assess the similarity between objects of different databases, we adopt four geographical and topological distances. From each one, a similarity measure is derived. We use:

- The **Hausdorff distance** (Rucklidge, 1996). It is a straightforward distance between two subsets of a metric space (the lines in our case). We did not use the Fréchet distance, which is most suited for curves, as the Hausdorff measure is widely applied to linear object matching. Also, 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:

$$d_{\text{Hausdorff}}(L_1, L_2) = \max\{\min_{a\in L_1, b\in L_2} d_{\text{Euclidean}}(a,b), \min_{a\in L_1, b\in L_2} d_{\text{Euclidean}}(b,a)\} \tag{2.15}$$

  Where a and b are the end nodes of L_1 and L_2.

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

$$d_{\text{length}}(L_1, L_2) = \frac{|\text{length}(L_1) - \text{length}(L_2)|}{\max(\text{length}(L_1), \text{length}(L_2))} \tag{2.16}$$

- The **Orientation**-based distance. The corresponding pipes should also have similar orientations, and a strong deviation between lines may indicates 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 defined for a node N as the number of lines connected to N. We use this distance to check whether the objects to match are having the same neighbourhood structure. Depending on the rate of missing nodes, this measure may be misleading. The distance in terms of the node degree between two nodes n_1 and n_2 is computed as follows:

$$d_{\text{degree}}(n_1, n_2) = \frac{|\text{degree}(n_1) - \text{degree}(n_2)|}{\max(\text{degree}(n_1), \text{degree}(n_2))} \tag{2.17}$$

  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 (Figure 17). The mean of the degree-based distance of the two couples is retained as the node degree-based distance between the two lines:

$$d_{\text{degree}}(L_{a,b}, L_{c,d}) = \frac{d_{\text{degree}}(a,c) + d_{\text{degree}}(b,d)}{2}, \text{where} \quad d_{\text{E}}(a,c) <= d_{\text{E}}(a,d) \tag{2.18}$$

  Where $d_E(.)$ is the Euclidean distance. Computing end node degrees independently avoids situations such as the one in Figure 17, where the sum of the node degrees of line $L_{a,b}$ is equal to the node degrees of line $L_{c,d}$ but where the relationships are not semantically identical, since all the connections are in the node c.

We divide the selected distances into two categories. The first one includes Hausdorff, length and orientation-based distances whereas the node degree forms the second category. Indeed, corresponding pipes from two sources are generally close and have similar lengths and orientations. However, when data sources have significant and non-uniform discrepancies and distortions, the objects which correspond are not necessarily the closest in terms of Hausdorff, length- or orientation-based distances. These three distances suggest non-corresponding couples when their values are high.

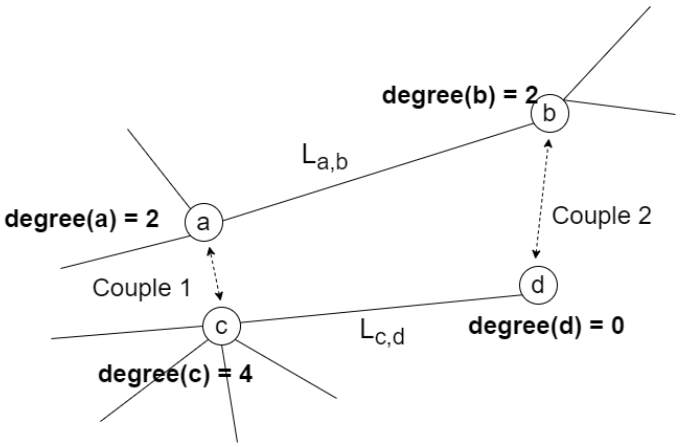


Figure 17 Example of the node degrees of two lines.

They cannot confirm peremptorily the correspondence when small, but indicate potential matching couples. On the contrary, a high value of the node degree distance could be due to data incompleteness, so that the non-matching cannot be confirmed. In this case, the small values are the ones that indicate potential matches. This is particularity true when strokes are considered as corresponding units since having different candidates with the same node degrees is not frequent.

As indicated in paragraph 2.3.2, one of the main steps in the process of using the DS theory for the matching objects consists in transforming distances into mass functions. Accordingly, this distinction in interpretation must be considered in this step. In paragraph 2.4.2.2, we show how these differences are modelled in our new process.

2.4.2 The proposed approach for matching wastewater spatial objects

2.4.2.1 Stroke/line based approach

Our process of matching lines representing wastewater networks is illustrated in Figure 18. As a first step, we propose to use strokes as a matching unit to overcome the constraint of missing nodes. Later on, we use pipes as a matching unit for the remaining lines that do not belong to any matched stroke. In stroke detection algorithm (see algorithm 1), we consider two connected pipes as part of the same stroke when their angles differ by less than 20 degrees. Indeed, a lower value will consider that a very small set of pipes is part of a stroke, whereas a too high value will lead to complex shapes unlikely to be matched. We set the value to 20 as a compromise between these two facets. We repeat this operation until all the lines are processed.

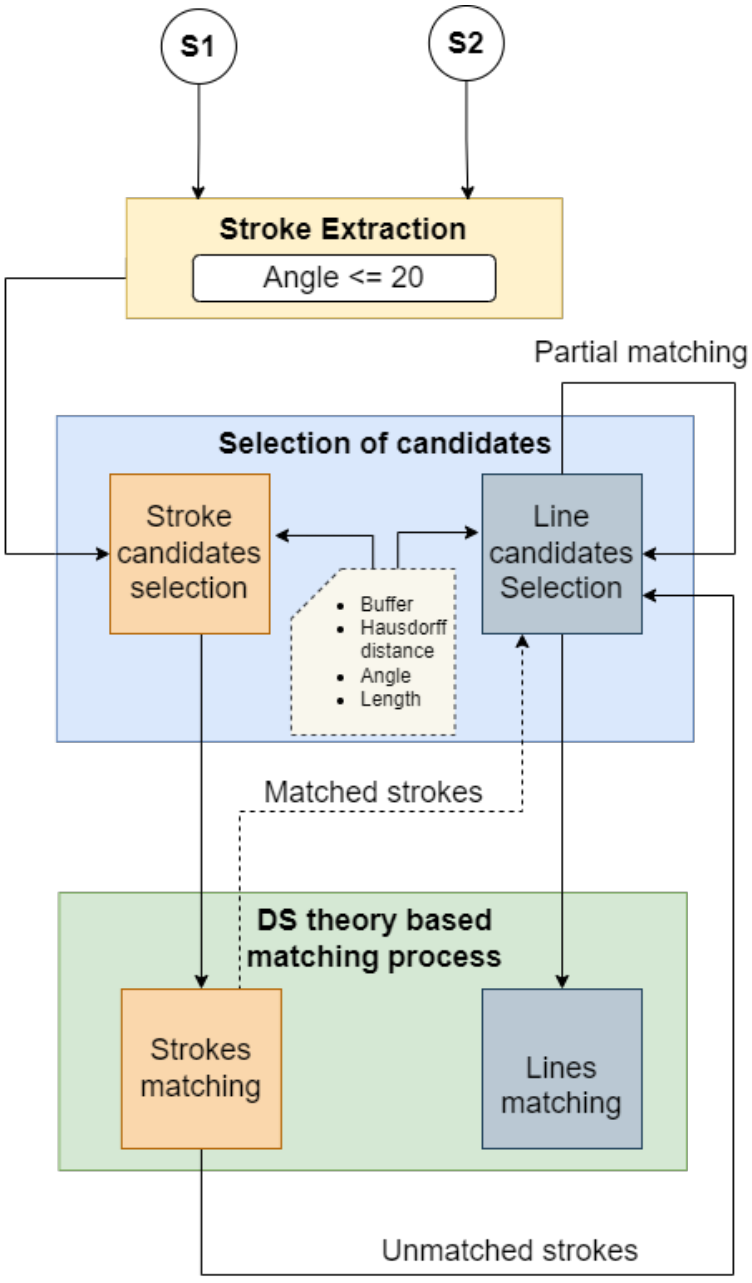


Figure 18 The proposed process for matching wastewater networks.

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
```

In the second step of the matching process, the purpose is to select, for each stroke of reference, the potential candidates to be analysed by the DS theory to eventually be able to take a matching decision (see Figure 18). The selection is performed by using a fixed size buffer as well as distance filters to eliminate unnecessary candidates. In our application, the buffer's width is set to 15 meters, which is large enough to have no impact on the results as candidates will be excluded further on by the filters. The buffer is only applied to reduce the computation time of distances. Afterwards, the candidates are disqualified when one of the following constraints is not satisfied:

- $d_{\text{Hausdorff}} \leq 40\text{meters}$.
- $|\theta_{L_1} - \theta_{L_2}| \leq 45$.
- $d_{\text{length}} \geq 0.8$.

Indeed, we consider that beyond a distance of 40m and a difference in orientation of 45°, the objects are unlikely to correspond. Moreover, if the length of a candidate is less than 80% of the length of the reference object, a partial matching should be conducted.

In the third step, we use the DS theory to combine the similarity measures based on the distances defined above (see section 2.4.1). This step is detailed in the next paragraph where we match the strokes and compute the uncertainty of the matched couples based on the plausibility and pignistic probability. An optional step can be added to match the pipes forming the matched strokes. At this stage, methods such as in (Walter & Fritsch, 1999), can be applied to achieve pipe matchings of cardinalities

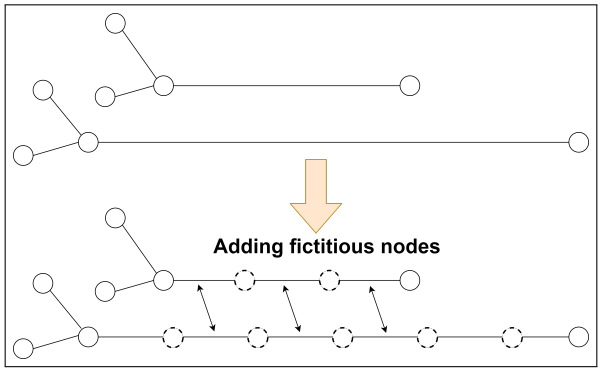


Figure 19 Example of adding fictitious nodes to lines representing the pipes in order to achieve partial matching.

1:1, 1:n and n:m. For the remaining unmatched strokes, we use the pipes directly as matching units. Like for strokes, the similarity measures' combination and decision are based on the DS theory.

As for the partial matching, and since wastewater networks are often expanded and repaired, the remaining unmatched lines include not only the pipes which actually don't have a corresponding candidate, but also misrepresented and newly laid pipes. We propose to handle situations like the one illustrated in Figure 19, 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.

2.4.2.2 Enhanced DS theory based process

In the 3rd step, we use the DS theory to combine the distances and decide whether two or more objects match. We enhanced the DS based process introduced in (Appriou, 1991) and applied in (Nassreddine et al., 2009; Raimond et al., 2015) (see paragraph 2.3.2). Considering two independent sources including N and M objects respectively, we define the frames of discernment for the local measures' combination as $\Omega_1 = \{C_{i,j}, \neg C_{i,j}, E\}$, where $C_{i,j}$ is the mass value associated to the potential corresponding objects indexed by $i \in \{1 \dots N\}$ and $j \in \{1 \dots M\}$. $E = \{C_{1,1} \dots C_{N,M}\}$ is the frame of discernment of the candidates combination step, of size $N \times M$.

Figure 20, illustrates our three core contributions. Compared to the original process (Figure 16), we modified the two phases of the transformation step by taking into consideration the candidates' ranking and mixed models. We also added a combination

step that we called bidirectional measure combination after the mass initialisation. Our three propositions to enhance the original model are detailed in the following:

- Candidates' ranking: the first phase of the transformation step consists in transforming distances into similarity measures. Since one object can have several very close candidates, ranking candidates in terms of each considered distance is important. This piece of information is relevant and should be exploited in the matching process in order to highlight the closest candidates. This also avoids that similar distances become contradictory masses and generate an important conflict after the combination step. Candidates' ranking has never been considered in the process of the DS theory. In the new process, we propose to rank the candidates in the first transformation step using eq. 2.19, inspired by the work described in paragraph 2.2.1 eq. 2.1 (Beeri et al., 2004).

$$\text{Similarity}_{\text{metric}}(C_{i,j}) = \frac{d_{\text{metric}}(C_{i,j})^{-\beta}}{\sum_{k=1}^{\text{N}_\text{c}} d_{\text{metric}}(C_{i,k})^{-\beta}} \tag{2.19}$$

  where N_c is the number of candidates, d_{metric} is a distance such as Hausdorff, β is a rate to decrease the measure when the distance increases compared to the other potential candidates $k \in 1 \ldots |\text{candidates}|$. 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 term of the Hausdorff distance.

- Mixed model: the second phase of the transformation step aims to transform the similarity measures into masses. As described above, (Appriou, 1998) proposed two different models to assign masses to each focal element. Contrary to previous studies (Deng et al., 2019; Nassreddine et al., 2009; Raimond et al., 2015), where only one of the models is applied, we propose to use both models (mixed models) depending on the measure to transform (see paragraph 2.3.2). We consider that the measures based respectively on the Hausdorff distance, the length and the orientation can be modelled by model 1. Indeed, when the values of these measures are null 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 they provide only an indication among others. We chose to use model 2 for the node degree measure,

since the connectivity of the strokes or the pipes, when similar, provides an important evidence that these candidates match.

- Bidirectional measure combination: initialization of the mass functions is usually carried out based on the computed distances from a reference set to a comparison set (one direction). In our approach, we consider the ranking of the candidates when defining the similarity measures. Thus, unlike previous studies where the local combination is conducted after the mass being initialized (Figure 16), we propose to combine the masses first in both directions of the sources before combining them with the rest of the measures. This is carried out in order to deal with situations where the candidates' ranking vary based on the direction of the matching.

The local measures' combination and the candidates' combination steps were not modified and all the combination steps are computed using the normalized conjunction operator. We choose to use the conjunctive rule rather than the disjunctive one which is less precise as it generates larger sets after the combination. We choose the matching objects based on the plausibility value.

2.5. Experiments and real case dataset

In order to validate our approach, we first performed tests on synthetic data, then on real-world datasets. Indeed, we applied the process described in Figure 20 on synthetic data with the aim of proving the effect of each of our choices, that is the transformation of distances to similarity measures while considering the ranking of the candidates, the transformation of similarity measures to mass functions by selecting the suitable model and the bidirectional combination. We compared these choices to those of (Raimond et al., 2015). As for real-world datasets, since we do not use semantic attributes as in (Raimond et al., 2015) and we conduct the matching first using the stroke concept, comparing the performance to other approaches would be irrelevant, especially since no standard datasets of wastewater networks are available. The performance of our new proposition is assessed using real data at each phase of the process (Figure 18).

Five configurations, summarized in table 2, were defined for the experiments:

- Configuration 1: the similarity measures are initialized without taking into consideration the candidates' ranking according to the distance metrics. We use the initialization method introduced in (Raimond et al., 2015) as described in paragraph 2.3.2. The transformation of the similarity measures to mass functions is carried out using model 2 (equation 2.14).

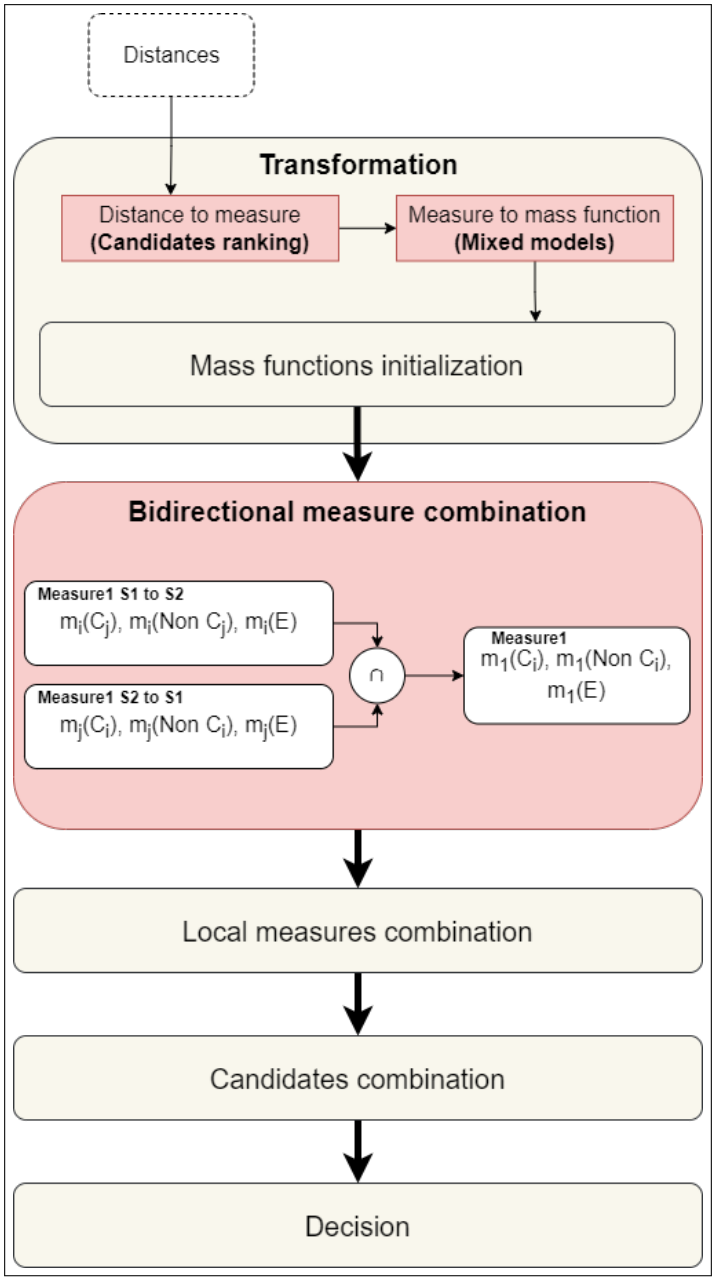


Figure 20 Our enhanced DS theory process for matching wastewater networks.

Table 2 Summary of the configurations.

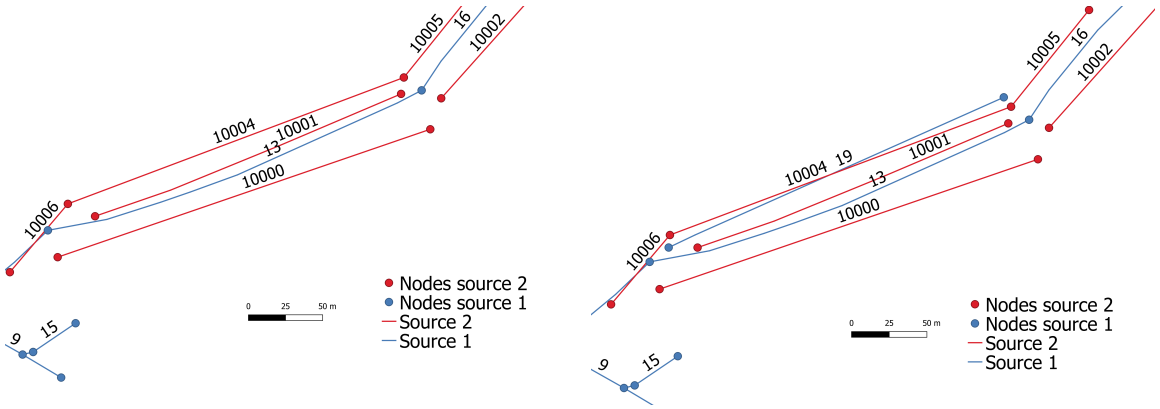
	Config.1	Config.2	Config.3	Config.4	Config.5
Ranking			X	X	X
Model 1		X		X	X
Model 2	X		X		X
Bidirectional					X

- Configuration 2: the similarity measures are initialized as in configuration 1. However, the transformation of the similarity measures to masses is based on model 1 (equation 2.12).
- Configuration 3: the similarity measures are initialized with respect to the ranking according to the distance metrics. The masses are defined with respect to model 2.
- Configuration 4: the similarity measures are initialized as in configuration 3, but the masses are defined with respect to model 1.
- Configuration 5 "mixed"): the similarity measures are initialized as in configuration 3 and 4. The mass values are defined with respect to model 1 and model 2, depending on the measures. As described in paragraph 2.4.2.2, we use model 2 for the similarity measures based on Hausdorff distance, orientation and length. Model 1 is used only for the node degree measure.

To efficiently demonstrate our choices on synthetic data, we first apply the Hausdorff distance only, in one direction of matching, without considering the bidirectional measure combination and local measures' combination steps. Later, the length, the orientation and the node degree based measures are also considered along with the bidirectional step. At this stage, we compare our mixed model proposition to configuration 4. For real-world datasets the results are described and discussed based on the configuration mixed only. In addition since we consider that the data sources are not always reliable, due to distortions, vectorisation errors and missing data, we apply a discounting based on the source's reliability.

2.5.1 Synthetic data

Figure 21 shows lines delimited by their associated nodes from two sources (source 1 in blue and source 2 in red). We considered two use cases in 21a and 21b. Table 3 shows the corresponding Hausdorff distance in meters between the lines of the two sources.



(a) Use case 1. (b) Use case 2.

Figure 21 Synthetic use cases.

Table 3 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

The two use cases differs by the presence of the object identified as '19' 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 by '10000', '10001' and '10004'. Based on Figure 21a, taking into consideration the Hausdorff distance (table 3) and the node degrees of the end nodes of the lines, the corresponding objects should be the couples ('13','10004') and ('19','10001').

2.5.2 Real-world datasets

We consider two real-world datasets from Prades-le-Lez (Figure 22), a small city of 5 908 inhabitants (INSEE, 2019) 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 23. Second, although the 2017 network is more recent and contains more pipes, we can see that both datasets have missing pipes. Third, the nodes which 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). We compare our matching results to the matching achieved manually by the operator. We use the precision and the recall metrics to evaluate our process as follows:

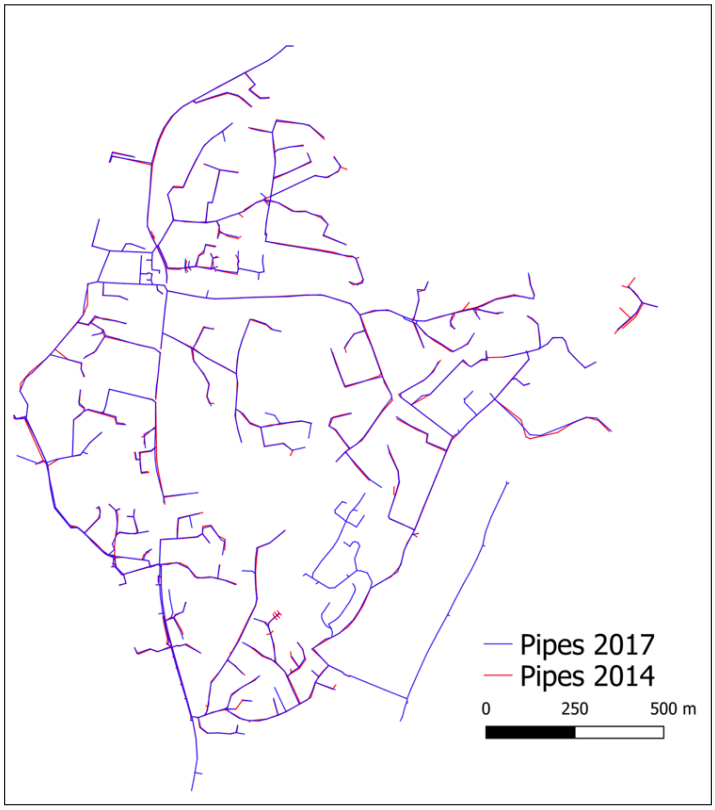


Figure 22 Maps of the wastewater network in 2014 and 2017.

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}} \tag{2.20}$$

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}} \tag{2.21}$$

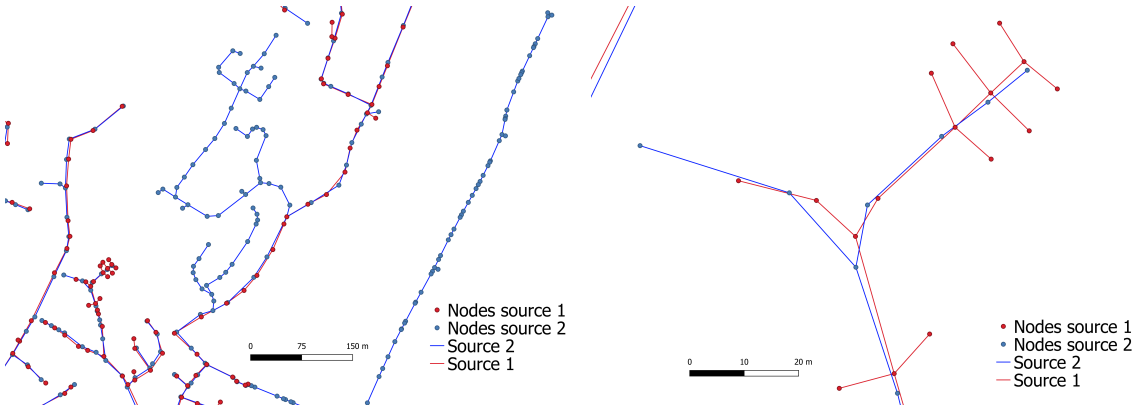
2.6. Results and discussion

In this section, we present and comment on the results obtained with our method, using both synthetic and real-world data.

2.6.1 Results using synthetic data

2.6.1.1 Use case 1

In use case 1, the reference object is identified by '13' in source 1 and the candidates are '10000', '10001' and '10004' in source 2. The mass values ($m(C_{i,j})$) for the four configurations are initialized as illustrated in Figure 24, based on the Hausdorff distance between lines of sources 1 and 2 (reported in table 3). The Hausdorff measure



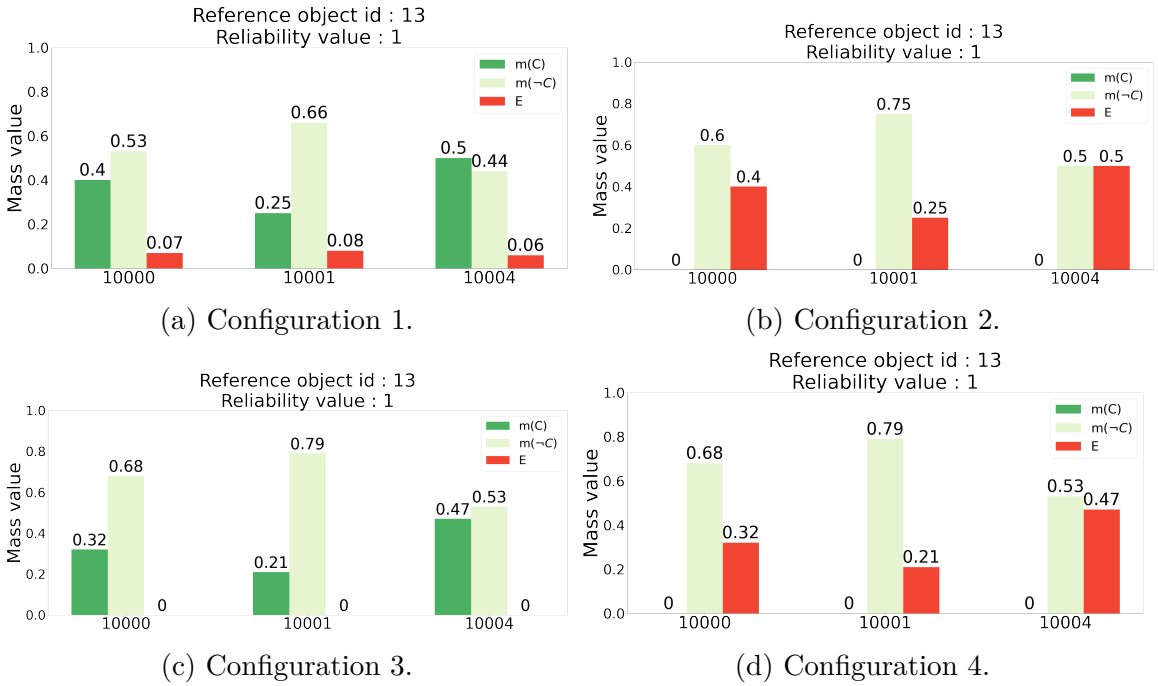
(a) Example of missing pipes in source 1. (b) Example of missing pipes in source 2.

Figure 23 Examples of data imperfections in Prades-Le-Lez's datsets.

is considered to be totally reliable (i.e. $r_i = 1$). For the first configuration, the mass values are initialized as indicated by equation 2.14 (Figure 24a). The value $m(C_{i,j})$ is also used with model 1 (equation 2.12) to initialize the mass values for configuration 2 (Figure 24b). As for configurations 3 and 4 (Figures 24c and 24d), we used equation (2.19) to initialize the mass $m(C_{i,j})$, and models 2 and 1 respectively. To evaluate the effect of the models on the conflict C generated by the combination of the masses, we used the conjunctive rule (equation 2.6), 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 25, shows the mass values after the candidates' combination step. We notice that for configurations 1 and 3, where model 2 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 26), which is often used as a decision criterion. However, the latter is 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 $E = \{('13', '10000'), ('13', '10001'), ('13', '10004')\}$ with only three possible couples.

2.6.1.2 Use case 2

Use case 2 (Figure 21b), is based on use case 1 with an additional object identified as '19' in source 1. The frame of discernment is then defined by E = {('13','10000'), ('13','10001'), ('13','10004'), ('19','10000'), '19','10001'), ('19','10004')}. Figure 27 shows the mass values' initialization for the reference object '19' in the four configurations. The masses obtained after the combination step are illustrated in Figure 28. 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 ranking of the candidates using equation (2.19), the mass value of



(a) Configuration 1.

(b) Configuration 2.

(c) Configuration 3.

(d) Configuration 4.

Figure 24 Initial mass values of the four configurations for the reference object '13'.

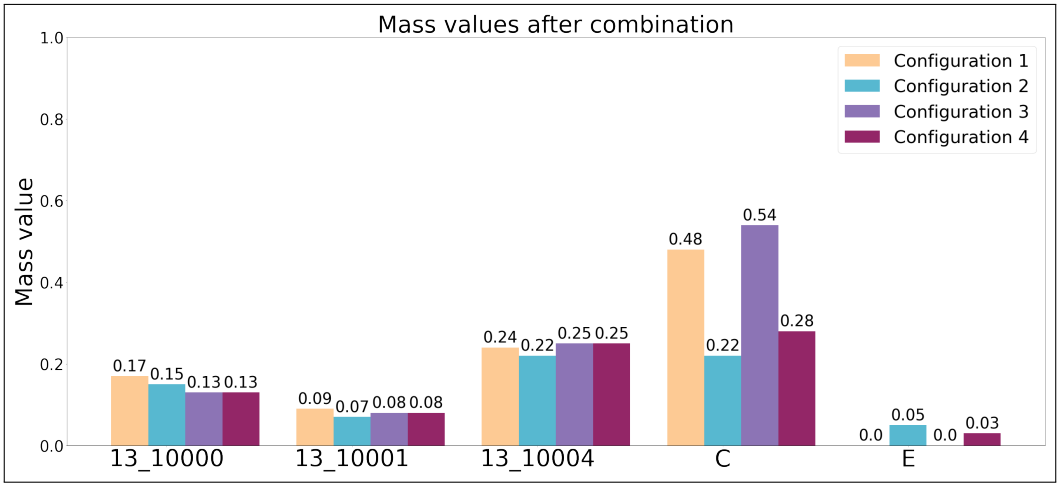


Figure 25 The mass values obtained after the combination process for use case 1.

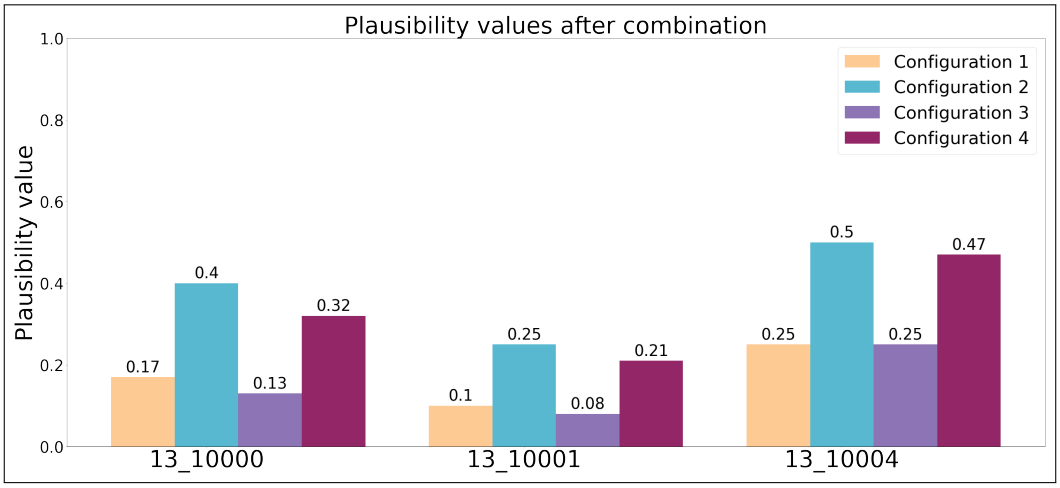


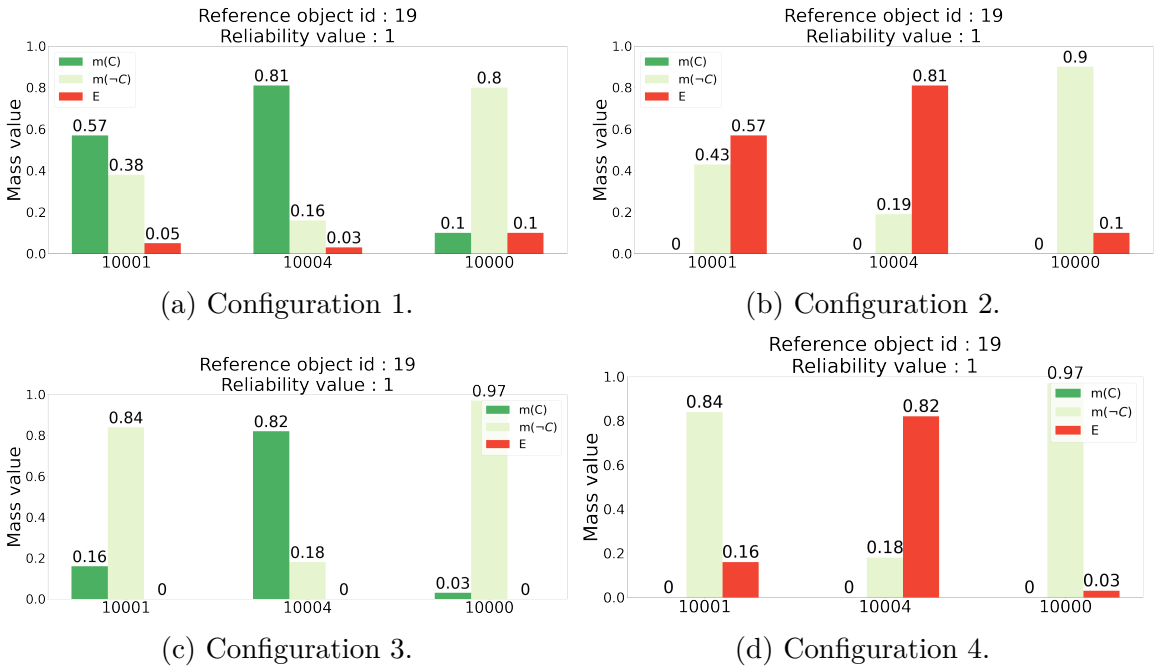
Figure 26 The plausibility of the corresponding couples after the combination process for use case 1.

the closest candidate for configurations 3 and 4 is almost twice that of configuration 1 and 2 (Figure 28). In addition, the plausibility values (Figure 29) are all almost equal to zero for configurations 1 and 3, in clear contrast to configurations 2 and 4.

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 a 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 29.

Using the Hausdorff distance only, the couples designated to match based on the plausibility values (Figure 29) are $(\prime 19\prime, \prime 10004\prime)$ and $(\prime 13\prime, \prime 10000\prime)$. This incorrect matching is to be expected, since only one measure is applied. In the following, we apply our process as described in Figure 20, and we compare configuration 4 to the mixed configuration (configuration 5).

Table 3, 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 the matching process by the bidirectional measure combination step. Figure 30b, shows the mass values after combination of both directions (Figures 27d and 30a) where the mass $m(\neg 13\_10004)$ is no longer equal to 0.53, as it was considered above (Figure 24d), but equal to 0.94, thus impacting the output values of the matching.



(a) Configuration 1.

(b) Configuration 2.

(c) Configuration 3.

(d) Configuration 4.

Figure 27 Initial mass values of the four configurations for the reference object '19'.

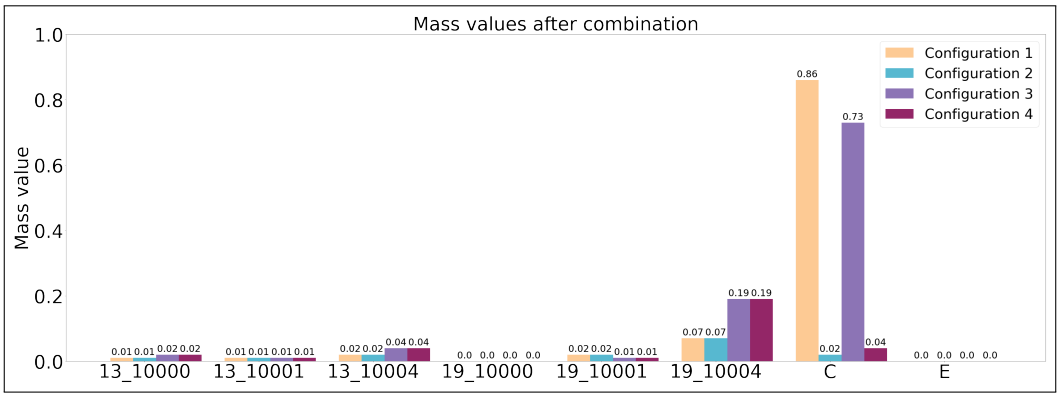


Figure 28 The mass values obtained after the combination process for use case 2.

Figure 31, 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, we obtain the correct corresponding couples $('13', 10004')$ and $('19', 10001')$. Hence, choosing the suitable model to combine information in the context of the DS theory can impact the final results greatly, while keeping the conflict value lower than with the use of model 2 only.

2.6.2 Results on real-world datasets

We followed our process described in Figure 18 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 yielded

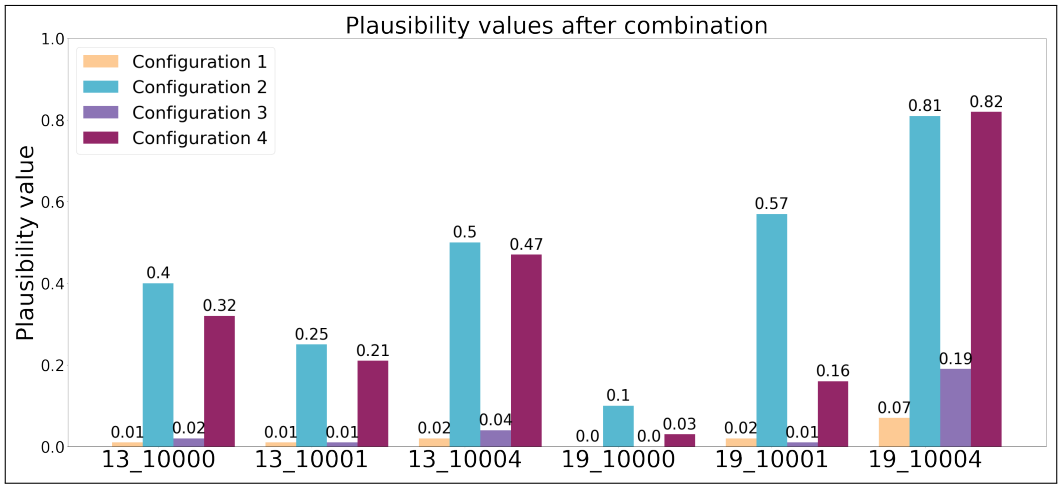
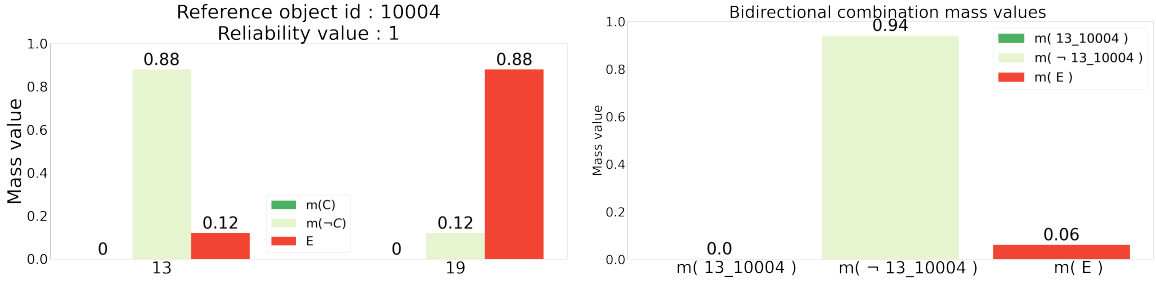


Figure 29 The plausibility of the corresponding couples after the combination process for use case 2.



(a) Mass values of the reference object identified by '10004' in source 2.

(b) Combination of the mass values of the couple ′13_10004′.

Figure 30 Bidirectional combination step of the Hausdorff-based mass values.

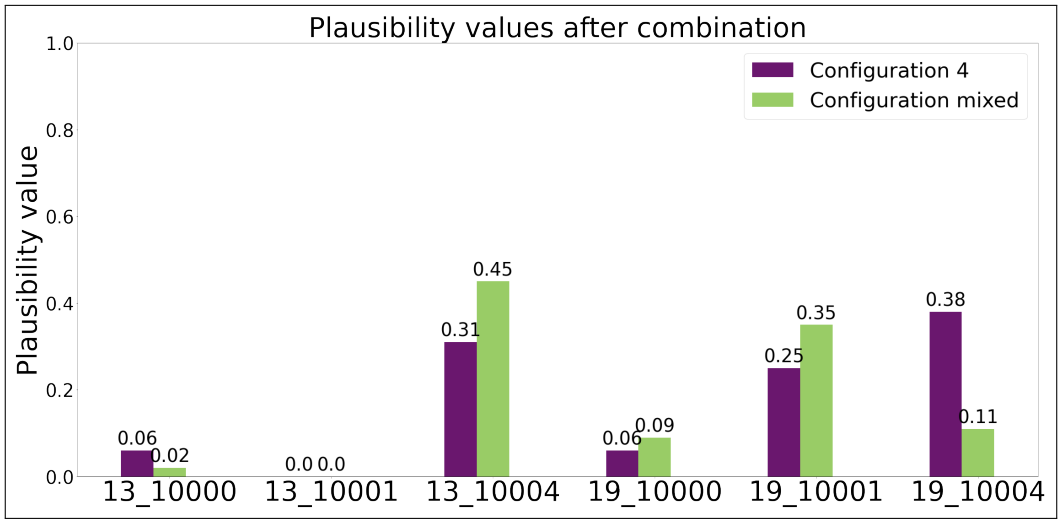


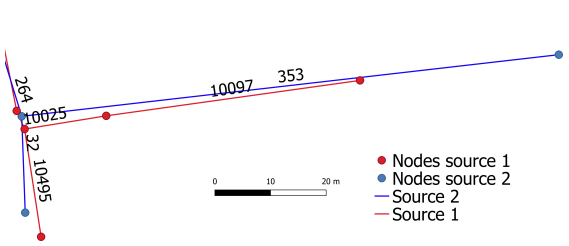
Figure 31 The plausibility values after applying our matching process for use case 2.

398 and 421 strokes respectively from source 1 and source 2. After extracting the strokes and identifying potential candidates for each stroke, our enhanced DS theory combination process (Figure 20) 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. In this phase, 277 strokes were matched. 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 these, 14 strokes were falsely matched, which gives a precision of 94.95%. No further operations were applied to the strokes that matched. However, an optional step of finding 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).

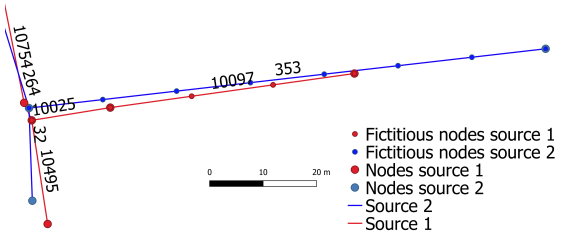
The same process is then applied directly on the remaining pipes, which did not match using strokes as matching unit. Considering missing nodes, we decreased the reliability of the node degree measure from 0.8 to 0.7. This yields the matching of 71 pipes from both sources, all being true positives. This 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 both sources.

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 32, shows an example of partial matching by adding fictitious nodes. As we can see in Figure 32a, pipes from source 1 identified by '10025' and '10097' can partially match with the pipe identified by '353' from source 2. This cannot by achieved in the previous steps, since nodes are missing. Figure 32b shows that no fictitious nodes were added to the pipe identified by '10025' since its length is smaller than 16 meters, whereas several nodes were added to pipes '353' and '10097'. Figure 32c 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.

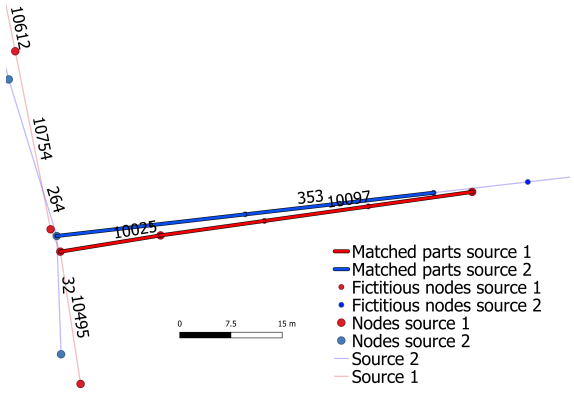
The results were compared the manual matching conducted by the operators to asses the quality of the matching. 731 pipes from source 1 should have a corresponding pipe in source 2. Using our proposition, we found 685 fully and partially true positives matched pipes from source 1. We counted 16 false positive matching, and 57 false negatives, that is a precision value equal to 97.7% and a recall of 92.3%. To keep



(a) Before adding fictitious nodes.



(b) After adding fictitious nodes 2.



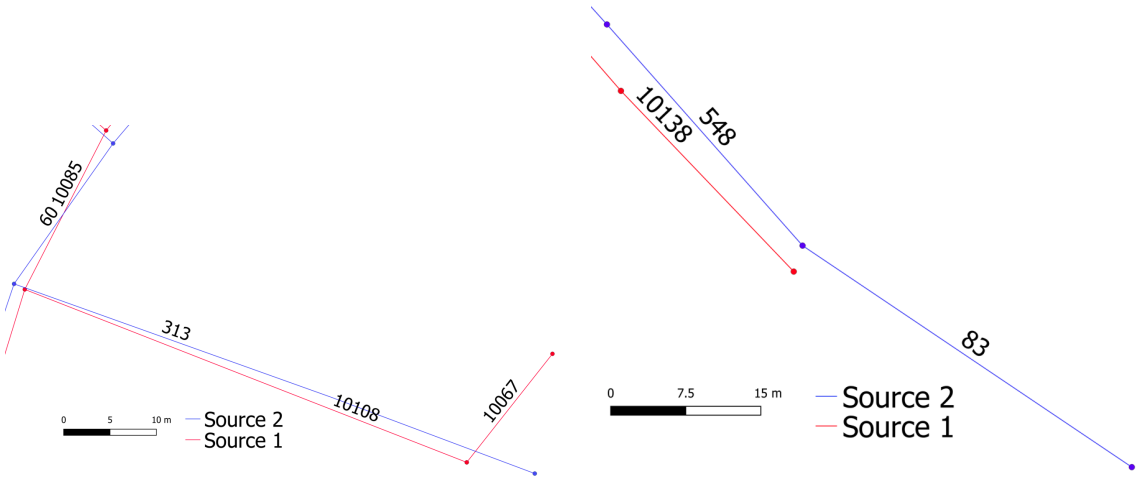
(c) Partial matching results.

Figure 32 Example of partial matching by adding fictitious nodes.

track of the uncertainty related to the matched couples, the pignistic probability and the plausibility values were preserved for each matching.

2.6.3 Discussion

In this study we 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 considered, only 55% pipes from source 1 are directly matched before the partial process compared to 80% of the pipes when the strokes are used. Also, when lines are directly used as matching units, the number of candidates could be high and several candidates for the same reference object may share similar geographic and topological structures. Consequently, as the frame of discernment grows exponentially with the number of candidates, the buffer and the thresholds used to limit the number of candidates must be carefully chosen. On the contrary, the candidates of a stroke are often limited, which makes line matching easier when carried out after stroke matching. Nevertheless, matching strokes must be conducted carefully, since the false matching of a stroke may result in the false matching of several lines.



(a) Example of missing pipes in source 1. (b) Example of missing pipes in source 2.

Figure 33 Examples of building a more complete dataset by adding missing pipes.

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 use case 1, no couple has a plausibility value greater than 0.6, which can be considered as an acceptable threshold. In this case, 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 object and a candidate have an attribute which 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.

The results of the matching can be exploited to confirm the presence of a pipe, to merge the attributes of the matched pipes or the create more complete datasets. Figure 33a shows an example where the strokes identified by '313' and '10108' are matched, and 10067 is missing in source 2. In a new dataset, pipe '10067' can be kept, since it is connected to pipe '10108'. Figure 33b shows a pipe identified by '83'

which can be recorded in the new dataset, since the pipes identified by '548' and '10138' matched.

2.7. Conclusion

In this work, we proposed a novel process for matching spatial objects, precisely wastewater networks. After identifying the origin of the imperfections related to wastewater databases and their consequences (long lifespan, multiple operators over the years and poor data management, which cause increases in costs and time of intervention) we described 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. To address the challenges due to missing nodes in wastewater databases and differences in temporalities and representation, in addition to using strokes as matching unit, we conducted a partial matching. We proposed a novel process based on the DS theory to combine similarity measures between the objects to match. The DS theory enables to model the uncertainty, the reliability and the imprecision at the input and the 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.

Although this matching process is intended for line objects of wastewater networks, it may be used for similar linear networks such as roads. The DS theory process is generic and can be adopted for matching nodes or polygons by defining the suitable distances. In addition, several operators are offered by the DS theory such as the refinement and the conditioning and can be used in future works to increase the level of detail after the matching operation, such as identifying the type of pipes.

In this study, we used only geographical and topological distances to achieve the matching of the pipes. We showed that distances based on attributes can also be applied even if they are available for a few objects. In addition, semantic information and regulatory rules related to the construction of wastewater network, such as the minimum length of a pipe, can be used to enhance the results and reduce the uncertainty of the matching.

3. Missing data imputation for wastewater networks

As indicated in the general introduction, both spatial information and attribute values suffer from incompleteness. As an answer to the third and the fourth scientific questions outlined in the general introduction, in this chapter we present our three contributions to address this issue:

1. *We start by proposing a set of algorithms based on domain knowledge and external information to impute missing attribute values required to run hydraulic simulations. We compare the imputed values to manually estimated ones. The results show a coherent hydraulic behavior. This work was presented at the 14th International Conference on Hydroinformatics.*
2. *Relying on Machine Learning approaches for data imputation, particularly Graph Neural Networks, we exploit the topology of the networks to predict missing attributes of wastewater networks. GNNs outperform classical methods especially when the rates of missing values are high. This work was published in "WATER" and is available here: <hal-03264611>.*
3. *In the third contribution we address the problem of missing spatial representation, namely pipes' spatial data. A new Wastewater Graph Neural Network (WaGNN) is introduced to predict pipe positions given manhole positions. The results show that our method outperforms popular GNN models in this task.*

3.1. Introduction

A problem often encountered when managing environmental systems, such as underground databases, is missing data (Junninen et al., 2004; Kofinas et al., 2018; Lin & Yuan, 2019; Schneider, 2001). In wastewater network databases, missing data may directly impact their management at both decision-making and business/scientific levels. Planning is an important task for decision-making. It helps develop a vision of needs in space and time, to quantify and prioritize them, to direct funding towards the most necessary investments and at a reasonable cost, since urgent and unexpected operation costs are far higher than the anticipated ones (ASTEE, 2015). Decision-makers use the available databases, which generally suffer from incompleteness, thus, often leading to delays in public works, traffic jams or collateral damage to the networks. Furthermore, experts in hydraulics who need to study the impact of external variables on the network, such as the discharge rate of subscribers into the network, use hydraulic modelling software, which require complete databases to run successfully.

However, few studies were published to help managers and the involved entities complete missing data. For instance, in (Chen & Cohn, 2011), the authors map underground networks using Bayesian fusion techniques to combine hypotheses extracted from Ground Penetrating Radar (GPR) with the spatial location of surveyed manholes and the expectations from the statutory records. Moreover, the authors in (Bilal et al., 2018) use a Bayesian mapping model to integrate knowledge extracted from sensors' raw data and available statutory records to infer underground network data including water pipe locations. To enhance the detection of underground networks, (Hafsi et al., 2017) fuse the data collected from different radars. In (Commandre et al., 2017), the authors apply deep neural networks to detect the position of manhole covers from high-resolution images. Although these propositions offer innovative methods to collect data, they are expensive and require long processing times and economic investments from the municipalities and the managers, which may not always be possible, especially for small towns.

Even if the network maps obtained through these procedures are in good agreement with the actual networks in terms of topology, they cannot be used directly by hydraulic modelling software. Indeed, no or very little information is available on the main attributes of the network: pipe shapes and dimensions, roughness, slopes, etc. A solution is then to resort to Missing Value Imputation (MVI) or Missing Data Imputation (MDI) algorithms, which try to replace the missing values of a data set to obtain a complete one. The goal is to estimate missing values based on the available ones. For instance, the authors in (Kabir et al., 2020) used MVI techniques to esti-

mate missing pipe diameter and age values, the number of service connections, and the number of valves. They mainly used statistical descriptors such as the distribution of attributes, the mean, the median, expectation-maximization, or the covariance matrix. Although the results were encouraging for some methods, this study had several limitations as the authors point out. For instance, in addition to being restricted to numerical attributes, this proposition was tested on a small percentage of missing attribute values with a maximum missing data percentage of 12.73% and a minimum percentage of 2.19% representing 63 pipes.

Many other studies address missing value imputation in different application domains. Their performances vary based on the type of targeted data: categorical, numerical, or mixed (Tsai & Chang, 2016), the percentage of missing data (García-Laencina et al., 2009) or the application domain of the completion task, such as biology (Liew et al., 2010) or pattern recognition (García-Laencina et al., 2010). MVI has been carried out using statistical techniques such as simple means, Multiple Linear Regressions (MLR), Logistic Regressions (LR), Random Forest Decision Trees (RFD) or, Bayesian inference (Bischof et al., 2018; Lin & Yuan, 2019; Murtojärvi et al., 2011; Ngouna et al., 2020; Serrano-Notivoli et al., 2017; Yadav & Roychoudhury, 2018). It now benefits from the most recent developments in Machine Learning techniques such as K-Nearest Neighbour (KNN), Support Vector Machines, Artificial Neural Networks, Long Short-Term Memory algorithms (Belda et al., 2020; García-Laencina et al., 2009; Giustarini et al., 2016; Ma et al., 2020; Nelwamondo et al., 2013) and more recently Graph Neural Networks (Spinelli et al., 2020). The latter are particularly interesting for missing value imputation on urban water networks whose design rules follow topological relationships both for network configuration and geometric properties.

3.2. Data imputation for wastewater hydraulic models

The present work aims at automatically completing the database associated to a network, using the little data available and rules based on classical guidelines for the construction of such networks. This contribution is organized as follows: Section 3.2.1 presents the methodology and the materials used in this study. The results are described and discussed in Section 3.2.2. Section 3.2.3 concludes this work.

3.2.1 Materials and methods

We assume that the map of a wastewater network is available in the form of a polyline shapefile which may be obtained either from the local stakeholder or from the previously developed mapping process. In this work, we use the wastewater network of Prades-le-Lez, a small town located in Southern France, produced in (Chahinian et al., 2019). We already know that no or little data is available in the attribute table associated to the underground network. Moreover, other information may be available from national or open access geographical databases. In France, they correspond to products developed by the French Geographical Institute IGN: Bd-Topo® for roads and buildings and RGE-alti 1m® for elevation.

The minimum information required to run a hydraulic software are: inlet positions, depth and associated input discharges, pipe geometries, slopes and roughness. Inlet positions are assumed to be known from the map. In a first approximation, pipe roughness can be assumed uniform given the most probable material used in the city. Pipe geometry can also be assumed circular with little impact on the results leaving one parameter to be estimated: their diameters. Finally, if the absolute depth of an inlet is not mandatory, the corresponding pipe's slope is a parameter of great importance, as classical hydraulic models cannot compute gravity fed flow on counter-slopes, unless pumping stations are added. We thus present the methodology used to automatically assign diameter and slope values to each pipe of the network and to assess input discharges.

Diameter estimation. Using minimal and maximal value bounds chosen by the user, the algorithm allocates the pipe diameters according to Strahler's order (Strahler, 1957), thus ensuring the general increase of the diameters from upstream to downstream. When a minimum of 20% of the attributes are available, one may use a semi-supervised learning method to predict the missing values as described in section 3.3.

Slope estimation The developed algorithm first allocates the ground elevation value minus the minimum burying depth of pipes to pipes' nodes, i.e. 0.8m in France. Starting from the outfall, it associates an upstream and downstream point, and thus a direction to each pipe, assuming gravity fed flow. However, even if ground elevation has proven to be well related to manhole depths, the precision of this simple estimation is not sufficient and requires modifications to ensure the conveyance of flow in the correct direction. The algorithm proposed here can be summed up as follows:

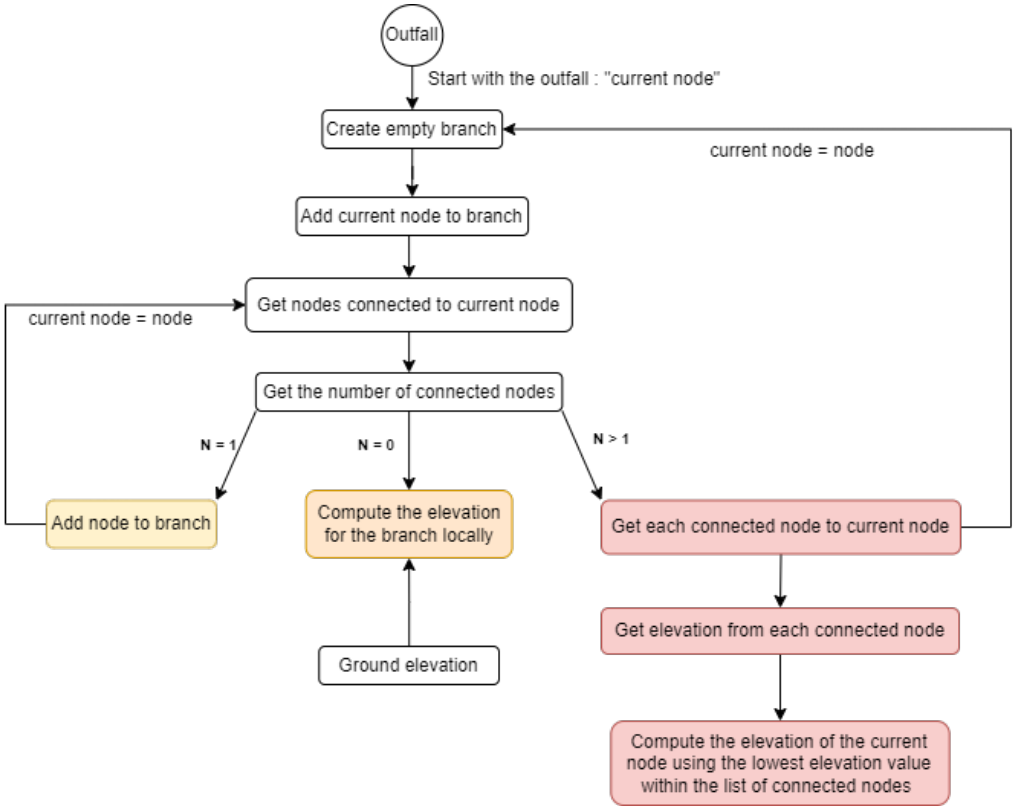


Figure 34 Slope estimation steps.

- Separate the network into "branches" where all the points can have one upstream and one downstream point at most;
- For each branch starting from the general outfall:
  - Check the coherence of the elevations and correct if needed to ensure a general decrease from upstream to downstream.
  - If several corrections are to be carried out on the same point, choose the one with minimum elevation.

Figure 34 describes the main steps of the slope estimation.

Input discharges At each node, input discharge is estimated from the following pieces of information: the average number of inhabitants in the surrounding houses and their mean daily sewage discharge assessed from drinking water consumption. We assume that each construction is more likely to be connected to the closest pipe in terms of Euclidean distance. Figure 35 presents an illustration of the process based on the IGN building database (BD-Topo©). Each Polygon is linked to the closest node and an input discharge is estimated for each node.

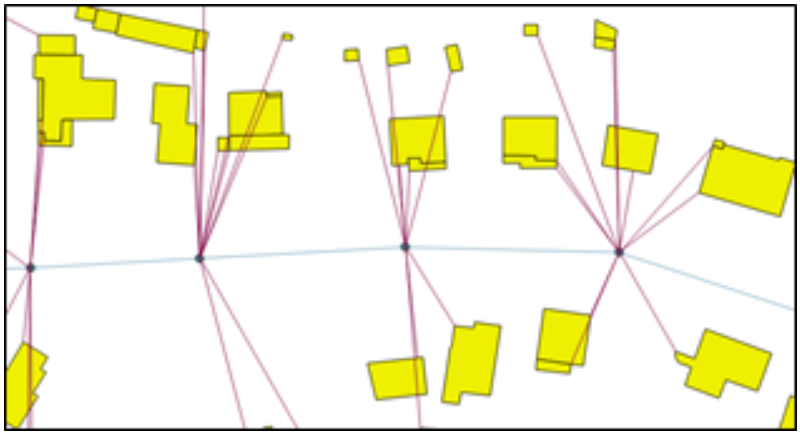


Figure 35 Linking each building to the closest network node.

SWMM© (Environmental Protection Agency, Storm Water Management Model, 2022) is a Windows-based desktop program, used for planning, analysis, and design related to stormwater runoff, combined and sanitary sewers, and other drainage systems. We use SWMM© to run the hydraulic simulation. The proposed process should result in the automatic creation of all the attributes required by this software. This will ensure considerable gain in preparation time for the user. Tests were carried on the wastewater network of Prades-le-Lez. The database contains approximately 800 pipes and almost no associated characteristics, namely no elevation.

3.2.2 Results and discussion

The procedure succeeded in completing the database and running hydraulic simulations without any errors. As no validation dataset is presently available for a better hydraulic validation, the attributes' values automatically estimated by the above-mentioned algorithms are compared with those manually estimated by a hydraulic engineer. For the elevation estimation the correlation is almost equal to 1 (Figure 36). Moreover, the simulation run on the entire dataset shows coherent hydraulic behavior (Figure 37).

Figure 38 shows an example of the hydraulic simulation using SWMM©, where the various colors indicate water flow in each pipe and manhole.

It should be reminded that the network used in this study is the result of the map created in (Chahinian et al., 2019), where mapping errors may occur and that our propositions to compute the missing values are carried out under the worst conditions i.e. where no attributes values are provided in advance. By running the hydraulic simulation, we were able to detect some mapping errors during this process, precisely on the slope estimation step.

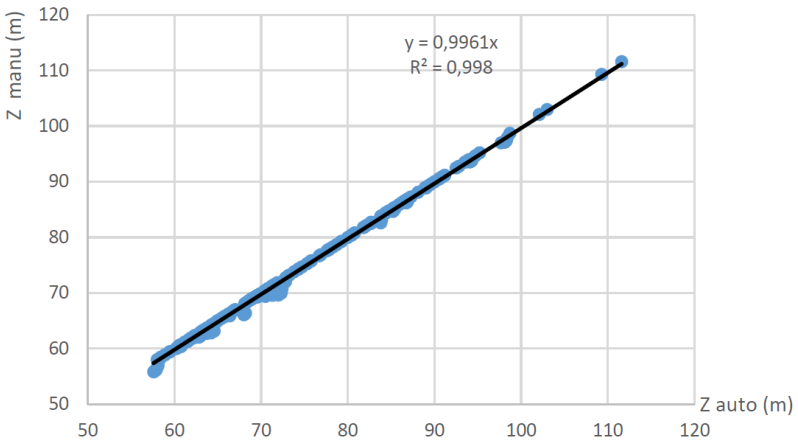


Figure 36 Automatic elevations (x-axis) and manual elevations (y-axis).

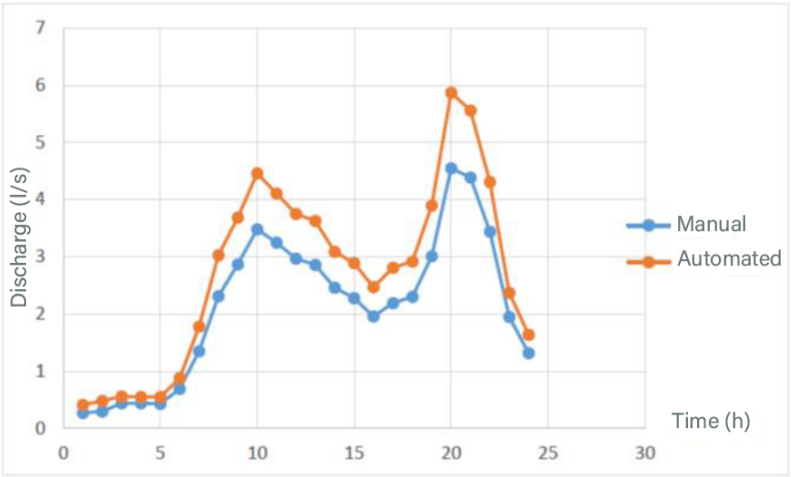


Figure 37 Comparison between output hydrographs from automatic and manual inputs.

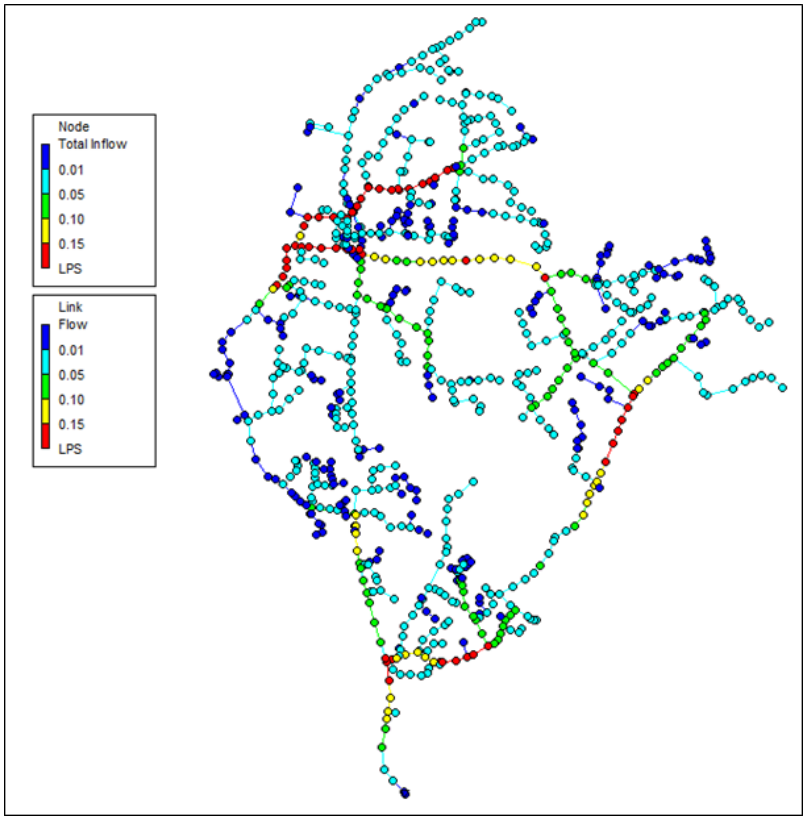


Figure 38 Simulation example using the SWMM© software with the proposed estimation process.

We noticed that assigning a slope value based on the elevation and the structure of the network may produce abnormal values such as a too large difference of altitudes between two nodes of the same pipe. Such errors impact the entire network, thus the simulation directly. We put forward a tracing algorithm to identify the pipes where the errors are initiated. After comparing the reconstructed map with the map provided by the operator, we noticed that all the pipes where the errors are reported are not part of the ground truth network and are due to detection or mapping errors. Figure 39 shows an example of such a situation, where the difference between the node identified by “657” and “738” is of almost 15 meters. The pipes from the reconstructed network (in blue) having the endpoints (“657”,”738”) and (“738”,”688”) do not exist in the actual network (in red).

3.2.3 Conclusions and perspectives for database completion

In this study we proposed a process to estimate the missing data needed to run hydraulic simulations using a different approach for each required parameter. We based our algorithms on information that are often available: the ground elevation, the average number of inhabitants per house and the mean daily discharge. In addition

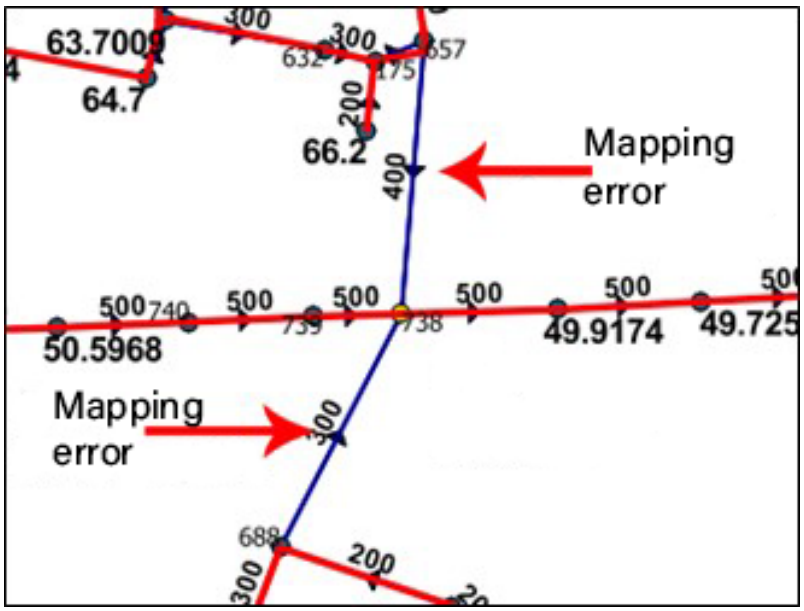


Figure 39 Identification of mapping errors.

to being quicker and easier to use, the results produced by the automated solution indicate a coherent hydraulic behavior when compared to manually estimated inputs.

The estimated data are not accurate and do not systematically reflect the reality on the ground. Consequently, the simulation results are inevitably uncertain and imprecise, and they can lead to wrong conclusions despite running without errors. In addition, estimation approaches are used when official recorded data are unavailable, and they are not intended to substitute for good quality field data. Thus, currently we are working on improving our process with the aim to account for the accuracy of the data that the user may define in the input map. Indeed, when the values of several pipe characteristics are known with certainty, they should be maintained by the algorithm and propagated to the surrounding pipes. For example, if a pipe's diameter is known, the algorithm should be able to modify other diameters to ensure the general increase of pipe conveyance from upstream to downstream. Furthermore, information about the networks can be collected from multiple sources using different techniques such as imagery and radars. Therefore, data fusion operations as described in Chapter 2, could help the managers surpass the problem of missing data and thus improve the accuracy of the simulations.

3.3. Data imputation using Graph Neural Network

The objective of this work is to use a Graph Neural Network to complete a wastewater network's database in view of hydraulic modelling of wastewater flow and helping

managers estimate the missing values in their databases. To the best of our knowledge, this is the first attempt to use MVI techniques based on machine learning techniques to infer the characteristics of a wastewater network. This contribution is structured as follows: Section 3.3.1 gives an overview of the approaches and methods of machine learning on graphs. Section 3.3.2 presents the methodology, the models, and materials used in this study. The tests and the results are described in Section 3.3.3. The conclusions and the discussion are in Section 3.3.4.

3.3.1 Background and state of the art

3.3.1.1 Machine learning and graphs

In the last decade, machine learning models, particularly neural networks, have been successfully used to accomplish a wide range of difficult tasks such as natural language processing (Collobert & Weston, 2008), image classification (He et al., 2016) and speech recognition (Graves et al., 2013). However, the models behind these achievements like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) are only adapted to Euclidean data and cannot be applied directly to graphs, as their structures may vary greatly from one graph to another. For example, CNNs widely used for image applications, exploit the fixed structure of the pixel's neighbourhood to define convolution filters with shared weights and pooling operators (Krizhevsky et al., 2017). This process cannot be directly generalised to graph structures since the number of neighbourhoods for each node might be different.

Considerable efforts have been deployed to make data represented by a graph benefit from the advancement of machine learning techniques. The main goal is to exploit the structure of graphs in the learning process, taking into consideration the topology and the relationships between their components (nodes and edges). Historically, machine learning models relied on handcrafted features, using approaches such as statistics to encode graph structures (Bengio et al., 2013; Zhou et al., 2019). For example, in the case of graphs used to model viewers' relationships, when the edges between nodes represent a common watched film, one may use the number of shared edges between two users to suggest new ones. However, these approaches are time-consuming and inefficient since they depend strongly on the type of application and the specific use case. To surpass these challenges, various automatic methods have been studied. Graph Embedding and Graph Neural Networks are the most common ones.

3.3.1.2 Graph Embedding

The goal of Graph Embedding is to use low-dimensional continuous vector representations for graph-structured data, instead of the whole graph, as input to the

machine learning algorithms. Graph Embedding is the overlap of two problems, graph analysis, which aims to extract useful information from graph data, and representation learning, whose goal, is to obtain a representation facilitating the extraction of useful information that is not necessarily low dimensional (Cai et al., 2018). Embedding techniques depend on the type of graphs used as input (such as homogeneous/heterogeneous, directed/undirected, etc.) and the type of desired output (nodes' embedding, edges' embedding, graph embedding). In (Cai et al., 2018) a clear taxonomy of the different techniques and applications of graph embedding is presented. Although graph embedding techniques have been successfully used in many applications such as node classification using the Node2Vec algorithm (Grover & Leskovec, 2016), they nevertheless present several drawbacks. Indeed, (Zhou et al., 2019) and (Cai et al., 2018) identified two severe ones: computation inefficiency and the inability to be generalised since they cannot deal with dynamic graphs. In addition, the authors in (Scarselli et al., 2008) indicate that mapping a graph structure into a simple representation may cause information loss. For example, in the case of node embedding, edges are considered as additional node features, although these links generally encode relationships between concepts or objects.

3.3.1.3 Graph Neural Networks

To operate directly on graphs, (Scarselli et al., 2008) proposed the first Graph Neural Network model. Described as the extension of existing neural network methods in the graph domain, this model considers nodes as concepts or objects and edges as relationships between them. To accomplish supervised learning, the GNN model associates each node to a state containing information about the node itself and its neighbourhood. Using a feedforward network, a shared transition function is defined to update all the states iteratively until a fixed point. The states are updated based on the current states of the nodes and those of their neighbours. Then, using a feedforward network, an output function is applied to the states to compute the outputs of each node, or a unique output for the whole graph, depending on the application. These steps are repeated using the descent-gradient algorithm until reaching the desired criterion. This GNN model has proven to be efficient in some application domains such as chemistry. However, it is not suitable for a variety of graph problems such as knowledge graphs and semi-supervised applications, where the goal is to predict missing data based on the graph structure. Besides, this model suffers essentially from the expensive cost of the computations while trying to reach fixed points. To address these problems, several variants of GNN models and new approaches have been proposed (Kipf & Welling, 2017; Y. Li et al., 2017; Thekumparampil et al., 2018). The most widely used one is the Graph Convolutional Network (GCN), which aims at generalizing CNNs to graphs. In the next paragraph, we present graph convolu-

tional network models for semi-supervised learning which might be used to complete missing data.

3.3.1.4 GCN for Semi-supervised learning

Graph Convolutional Network (GCN) models have achieved state of the art in many applications. In semi-supervised learning for node applications, the objective is to use labelled nodes to learn representations or embedding of both labelled and unlabelled nodes and therefore using the resulting representations to predict missing labels. GCNs are classified into two categories: spectral approaches and spatial approaches. Spectral approaches were first introduced in (Bruna et al., 2014). Since convolution filters, defined in the Euclidean space and used in CNNs, cannot be applied directly on graphs, (Bruna et al., 2014) have shown that they can be defined in Fourier domain for non-Euclidean data. This operation is defined in (Kipf & Welling, 2017; Zhou et al., 2019) as the multiplication of a signal $\mathbf{x} \in \mathbb{R}^N$ (one scalar for each node) with a filter $g_\theta = \mathrm{diag}(\theta)$ parametrized by $\theta \in \mathbb{R}^N$:

$$g_\theta \star \mathbf{x} = \mathbf{U} g_\theta \mathbf{U}^T \mathbf{x} \tag{3.1}$$

where $\mathbf{U}$ is the matrix of eigenvectors of the normalized graph Laplacian $\mathbf{L} = \mathbf{I_N} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, with a diagonal matrix of its eigenvalues $\mathbf{\Lambda}$. $\mathbf{D}$, $\mathbf{A}$ and $\mathbf{U}^T$ are respectively the degree matrix, the adjacency matrix of the graph and the graph Fourier transform of x. However, this proposition suffers from two major drawbacks. First, calculating the eigenvectors and eigendecomposition is computationally expensive, especially for large graphs. Second, the filters defined in the spectral domain are non-spatially localised, contrary to those in CNNs, *i.e.,* filters are not necessary applied to spatially close nodes. To surpass these challenges, improvements have been published, which generally consist in proposing new filters (Defferrard et al., 2017; Kipf & Welling, 2017). ChebNet (Defferrard et al., 2017) is the most popular one, and uses polynomial parametrization to compute K localised filters:

$$g_\theta(\mathbf{\Lambda}) = \sum_{k=0}^{K-1} \theta_k \mathbf{\Lambda}^k \tag{3.2}$$

where the parameter $\theta \in \mathbb{R}^K$ is a vector of Chebyshev coefficients. To address the computation issue, ChebNet uses Chebyshev expansion (Hammond et al., 2011) of order $K-1$ and $g_\theta(\mathbf{\Lambda})$ becomes:

$$g_\theta(\mathbf{\Lambda}) = \sum_{k=0}^{K-1} \theta_k T_k\left(\tilde{\mathbf{\Lambda}}\right) \tag{3.3}$$

where $T_k\left(\tilde{\mathbf{\Lambda}}\right) \in \mathbb{R}^{n\times n}$ is the Chebyshev polynomial of order k evaluated at $\tilde{\mathbf{\Lambda}} = 2\mathbf{\Lambda}/\lambda_{\max} - \mathbf{I_n}$, the rescaled eigenvalues in $[-1, 1]$ with $\lambda_{\max}$ the maximal eigenvalue. To alleviate the problem of overfitting on local neighbourhood structures on graphs, (Kipf & Welling, 2017) limit and simplify the filtering to only the first-order neighbours with $K = 1$.

Since they depend on the eigenbasis of the graph, spectral approaches cannot be used with graphs that have different structures. However, they are suitable for semi-supervised learning, which involves the prediction of features of the same graph used for the learning procedure. Thus, they are suitable for our goal, which involves the prediction of incomplete data related to wastewater networks.

Contrary to spectral approaches, spatial ones define convolution directly on graphs. Various propositions have been published: (Duvenaud et al., 2015) proposed a spatial convolution network that operates directly on graphs for molecular applications. GraphSAGE (Hamilton et al., 2017), one of the most popular frameworks in this category, defined as an inductive framework[2], is based on the aggregation of fixed-size node neighbourhood features:

$$\mathbf{h}^k_{\mathcal{N}(v)} \leftarrow \text{AGGREGATE}_k\left(\left\{\mathbf{h}^{k-1}_u, \forall u \in \mathcal{N}(v)\right\}\right) \tag{3.4a}$$

$$\mathbf{h}^k_v \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{CONCAT}\left(\mathbf{h}^{k-1}_v, \mathbf{h}^k_{\mathcal{N}(v)}\right)\right) \tag{3.4b}$$

where $\mathbf{h}^k$ denotes a node's representation at step k, $\mathcal{N}(v)$ the immediate neighbourhood of v, AGGREGATE is the aggregation function and σ a nonlinear activation function. Authors in (Hamilton et al., 2017) defined 3 aggregation functions: Mean, LSTM, and pooling. To avoid computing the spectrum of the graph Laplacian as in (Bruna et al., 2014) or (Defferrard et al., 2017) and to apply CNNs on graphs, (Du et al., 2017) proposed TAGCN, a method based on a fixed-size K-localized filters adaptive to the topology of graphs to replace the fixed square filters in traditional CNNs.

3.3.2 Materials and Methods

In this work, we seek to complete missing attribute values based on the structure of wastewater networks and the database records related to them.

[2] unlike transductive approaches that generate embedding for a specific seen fixed graph in their process, inductive ones generate low dimensional representation for unseen components of graphs

3.3.2.1 Models and test configurations

To highlight the added value of GCNs in this prediction task, we also apply algorithms that do not take into account topology. The GCNs' results will thus be benchmarked against these non-topological algorithms: Support Vector Machine (Cortes & Vapnik, 1995), Decision Trees (Safavian & Landgrebe, 1991), feedforward Artificial Neural Networks (ANN), precisely a MultiLayer Perceptron (MLP) (Rumelhart et al., 1986) and four GCN models that have proven to be efficient in many applications. The GCN models consist of two spectral models: GCN (Kipf & Welling, 2017) and ChebNet (Defferrard et al., 2017) as well as two spatial models: GraphSAGE (Hamilton et al., 2017) and TAGCN (Du et al., 2017).

Given that pipe diameters and materials directly impact hydraulic modelling results, which is the aim of our work, we chose to automatically predict the missing values for each one of these two attributes. Nevertheless, other attributes could be targeted the same way.

The available attributes and their missing values are not necessarily similar and vary between providers. Hence, to investigate whether GCNs are useful in real cases, we defined two configurations based on the available data:

- Configuration 1: The network graph, a portion of the values of the targeted attribute, and domain knowledge are provided.
- Configuration 2: The network graph, a portion of the values of the targeted attribute, domain knowledge, and other fields of the attribute table are provided.

When no attributes are available, domain knowledge can be used to create and add new attributes to the structure to improve the learning process. In wastewater networks, pipe diameters increase when moving from the upstream wastewater catchments to the vicinity of the treatment plant. This domain knowledge can be accounted for using Strahler's number, a measure of the network's branching complexity (Strahler, 1957). This attribute is easily computed for each pipe since the position of treatment plants is usually known. Thus, the first configuration is conducted using the network graph and Strahler's number as a domain knowledge attribute.

In the second configuration, managers possess more information about the networks, and relevant additional fields of the attribute table are used to infer relationships. Thus, this configuration is the richest in terms of learning material as it uses the network structure, domain knowledge, and additional characteristics to impute miss-

ing values. In this situation, the managers seek precise information about a specific attribute for various purposes, such as the diameter values for a hydraulic simulation.

For each of the two configurations, the datasets were split into two subsets: training and test. The training subset includes the available attributes of the pipes and their associated labels to be learned. However, contrary to non-topological models, in order to operate, GCN models require the structure of the graphs. Therefore, the entire structure of the graph modelled by the adjacency matrix of the wastewater network pipes was provided to this graph-based model. 10% of the training subset is used as validation subset to tune the models' parameters, that is the number of convolution layers, the number of epochs, etc.

For the MLP, we set the number of hidden layers to 3 with respectively 100, 50, and 25 units for the first, second, and third hidden layers. The number of outputs is defined by the number of classes depending on each attribute. The Rectified Linear Unit (ReLu) is used as an activation function between the layers. All layers are formed by the linear layers of PyTorch (Paszke et al., 2017) and the output is computed using the Log Softmax function. For GCN models (Figure 40) we set the number of convolution layers to 2, the number of hidden units was set to 20 for the first layer, and to the number of desired classes to predict for the second layer. Similarly to the MLP, we used the Rectified Linear Unit (ReLu) as an activation function between the two convolutional layers, and the LogSoftmax as the activation function to output the labels. For the ChebNet layers, the filter size K was varied from 10 to 40 depending on the configuration and the size of the training subset. For the SVM model, the regularization parameter C is set to 1, the radial basis function (RBF) is a degree 3 polynomial kernel function. For the DT models, the Splitter is set to "best", the quality of the split is evaluated by the "Gini" criterion without any max depth constraint.

We implemented the GCN models and the MLP using PyTorch (Paszke et al., 2017) and Pytorch Geomtric (Fey & Lenssen, 2019), where the name of the models GCN, ChebNet, GraphSAGE, and TAGCN are respectively GCNConv, ChebConv, SAGEConv, and TAGConv. The non-topological models: SVM and DT were implemented using Scikit-learn (Pedregosa et al., 2011).

3.3.2.2 Datasets

In this study, we used two real wastewater network databases. The first one is that of Angers Metropolis and is available through the French Government's open access portal[3]. The second source is the database of Montpellier Méditerranée Métropole

[3] https://www.data.gouv.fr/

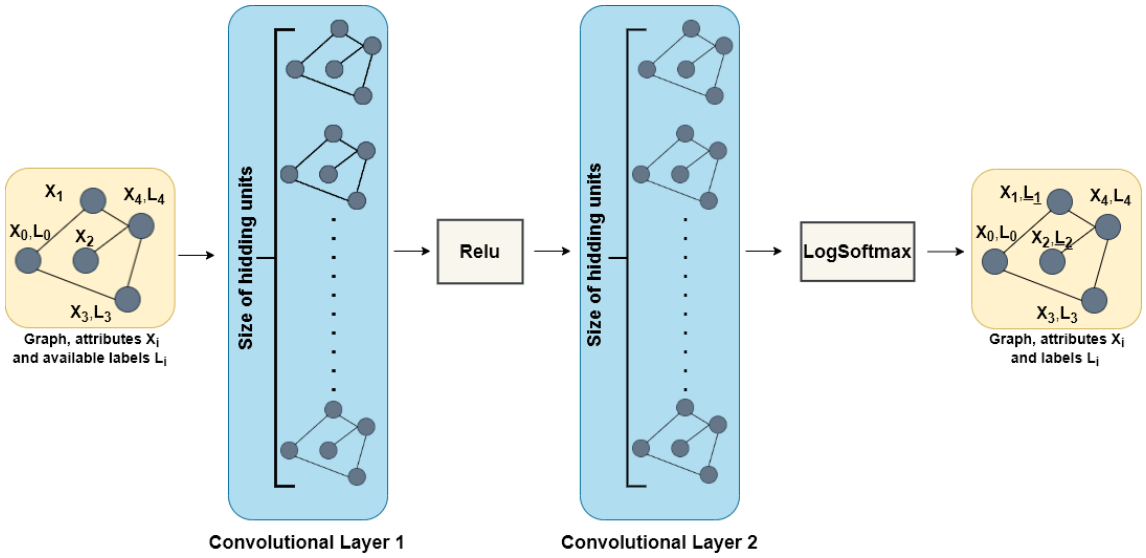


Figure 40 The Graph Convolutional Network models' architecture.

	datepose	exploit	ecounorm	longueur	materiau	gid	diametre
1	20031001000000	ANGERS LOIRE ...	GRAVITAIRE	75.35	PVC	4110	200
2	19691001000000	ANGERS LOIRE ...	GRAVITAIRE	36.25	AC	176	150
3	19500101000000	NON PRIS GEST...	GRAVITAIRE	28.28	AC	16566	200
4	19680801000000	ANGERS LOIRE ...	GRAVITAIRE	1.8	AC	13939	150
5	19950401000000	ANGERS LOIRE ...	GRAVITAIRE	64.12	PVC	12386	200
6	19500101000000	NON PRIS GEST...	GRAVITAIRE	18.28	PVC	25720	200

Figure 41 Example of an attribute table: Angers Metropolis in France ("Open platform for French public data", 2015).

(3M)[4]. We have chosen to use these databases since they have two specific fields for pipe diameter and material (see Figure 41 for an example of attribute tables). However, the attribute values are not all indicated and 5.9% of the total pipes of Angers and 28,63% of those of the Montpellier datasets have missing diameter or material values.

At the scale of a metropolis, wastewater networks are usually formed of several sub-networks of cities and villages, either managed separately or linked to the main treatment plant by a unique pipe. Thus, the acquired databases are composed of several sub-graphs that represent independent wastewater networks and Strahler's orders may be computed separately for each sub-graph. However, due to data imperfections, these disconnections may also be the result of missing spatial information such as missing pipes. Hence, to validate our results, this study was carried out on the sub-networks having the least missing attribute values. Taking into consideration possible spatial imperfections, we carefully extracted one sub-graph from each dataset (Figure 42):

- The Angers Metropolis sub-graph (Figure 42a) is composed of 754 pipes with only one unknown pipe diameter.
- The Montpellier Metropolis sub-graph (Figure 42b) is composed of 1239 pipes, with 44 pipes having unknown attribute values (either diameter or material).

The different materials encountered in Angers metropolis are Polyvinyl Chloride (PVC), Asbestos-Cement (AC), Cast Iron, and Metal. In Montpelier metropolis we found, PVC, AC, Cast Iron, Concrete, Glass Reinforced Plastic (GRP), and Polypropylene. Ten classes of possible diameters are present in Angers' subgraph and Montpellier's subgraph, ranging from 80 to 500. However, be it for materials or diameters, several classes have less than 10 elements and will not be considered in the following. Figure 43 shows the distribution of material and diameter attributes for the considered classes, for the two datasets.

3.3.2.3 Testing procedure

After tuning operations, the models are trained on 90% of the data and the remaining 10% are predicted. This is the first test. To put forward the models' ability to distinguish between classes and assess their effectiveness regarding minority classes,

[4] https://www.data.montpellier3m.fr/

(a) A sub-graph of Angers Metropolis wastewater network ("Open platform for French public data", 2015).

(b) A sub-graph of Montpellier Metropolis wastewater network (Montpellier Méditeranée Métropole, 2020).

Figure 42 Use case graphs.

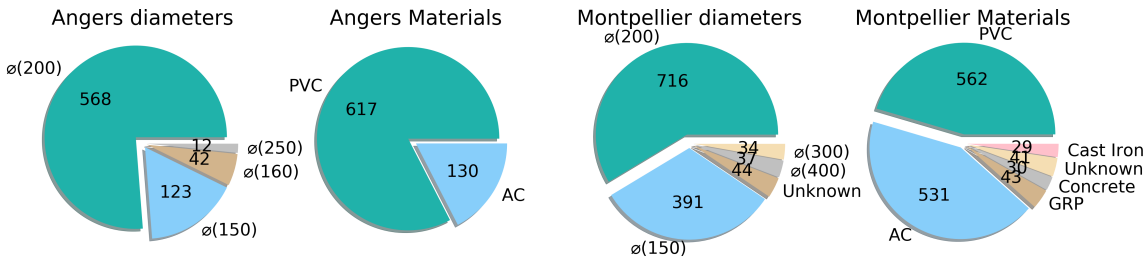


Figure 43 Diameter and material distribution for the Montpellier and Angers subsets. Only classes with more than 10 elements are represented here

we evaluate the results of the predictions by computing the Recall, Precision and F1-score metrics for each class of attributes such as defined in chapter 2, as follows:

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}} \tag{2.20}$$

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}} \tag{2.21}$$

$$\text{F1Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

This prediction operation is repeated ten times with randomly selected datasets to estimate the models' performance more accurately. The average of these predictions is examined. To evaluate the performance of the models over each attribute, we compute the Macro-Recall, the Macro-Precision, and the Macro-F1-score as follows, where N is the number of classes of an attribute:

$$\text{MacroRecall} = \frac{1}{N}\sum_{i}^{N}\text{Recall}_i \tag{3.6}$$

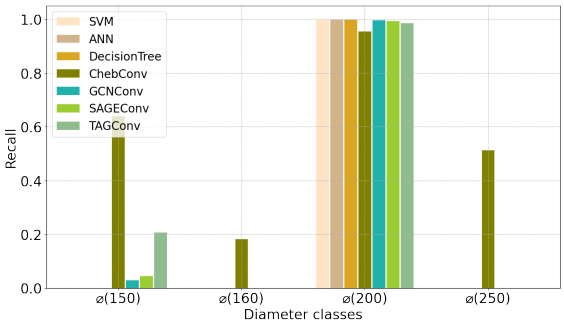
$$\text{MacroPrecision} = \frac{1}{N}\sum_{i}^{N}\text{Precision}_i \tag{3.7}$$

$$\text{MacroF1Score} = \frac{1}{N}\sum_{i}^{N}\text{F1}_i \tag{3.8}$$

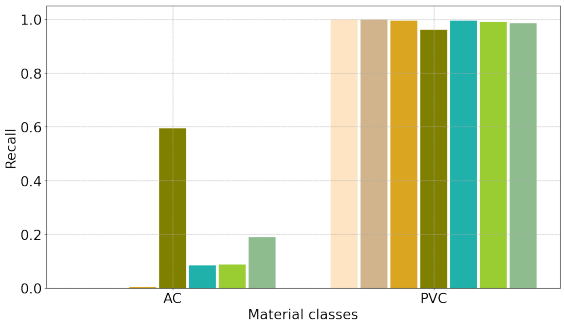
The training set is then sequentially reduced to increase the size of the test set, i.e., 80% for training and 20% for testing and so forth. As shown in Figure 43, attribute values are unbalanced, and the portion of the selected test subset may include only the dominant classes. Therefore, the test subset is extracted as a portion of the number of occurrences in each class. Consequently, only classes with more than 10 occurrences are considered as test subsets. For example, the diameter class of value $\phi(200)$ having 568 occurrences in the sub-graph of Angers metropolis, the number of selected pipes for a 10% testing subset (when the task is to predict pipe diameter values) will be 56.

3.3.3 Experimental results

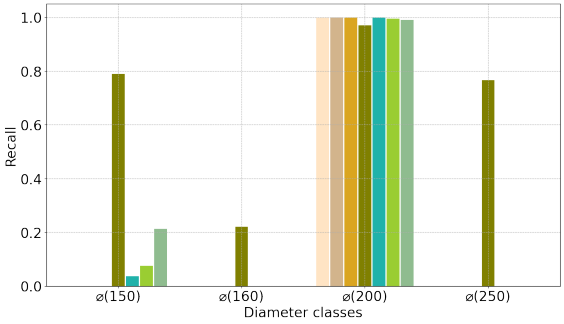
In this section, we show the results of the prediction of the attributes "Diameter" and "Material" for the two configurations described in section 3.3.2.1. We compare the results of several experiments using the different machine learning techniques presented in the previous section. The purpose of comparing GCN based algorithms



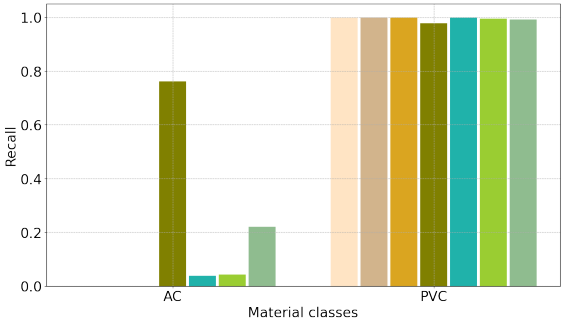
(a) Diameter prediction: 30% training and 70% test.



(b) Material prediction: 30% training and 70% test.



(c) Diameter prediction: 70% training and 30% test.



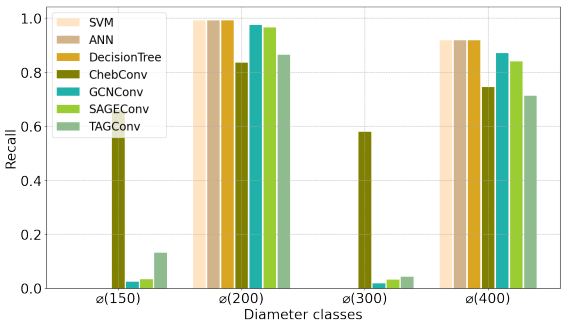
(d) Material prediction: 70% training and 30% test.

Figure 44 Configuration 1: Diameter and Material prediction for the Angers dataset for each class of the two attributes, evaluated using Recall score.

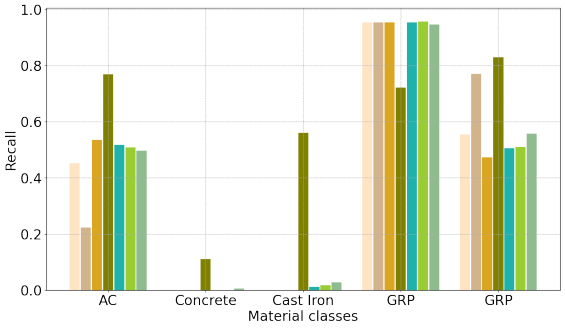
with different techniques of machine learning which do not use the graph's structure to predict missing data, is to investigate whether the network graph can facilitate missing data completion in the context of a machine learning approach. It is important to note that thanks to its structure, a GCN can predict classes without being given any attributes as input. This is clearly not possible for non-topological models. Thus, before conducting the experiments on the two defined configurations, and in order to see the behaviour of a GCN in terms of the quality of its results using only the structure of the graphs, we tested this possibility. The results show that GCN models GCNConv, SAGEConv, and TAGConv predict only the dominant classes, but the ChebConv model can identify other non-dominant classes albeit with very low recall scores such as 10% for the diameter class $\phi(150)$ on limited randomly selected test datasets. The prediction of minority classes with ChebConv, even with low scores, shows that using the structure of wastewater networks is promising.

3.3.3.1 Configuration 1

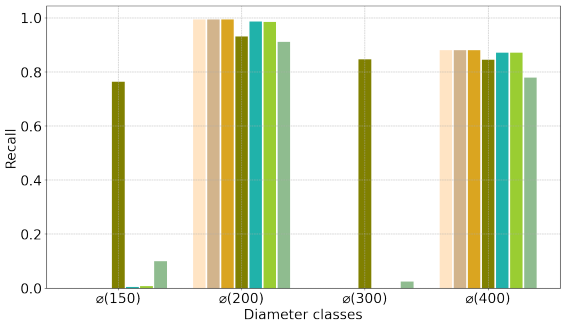
In addition to the portion of the available values and the structure of the network, in this configuration, we added Strahler's order as an attribute to help the models distinguish between the classes.



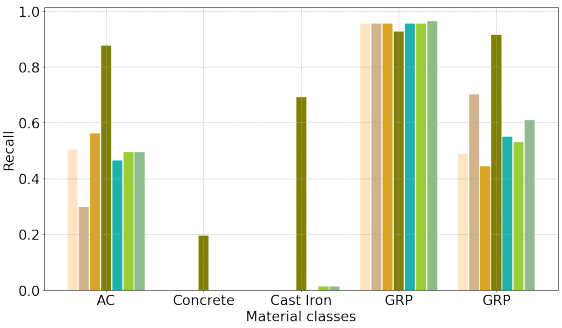
(a) Diameter prediction: 30% training and 70% test.



(b) Material prediction: 30% training and 70% test.



(c) Diameter prediction: 70% training and 30% test.



(d) Material prediction: 70% training and 30% test.

Figure 45 Configuration 1: Diameter and Material prediction for the Montpellier dataset for each class of the two attributes, evaluated using Recall score.

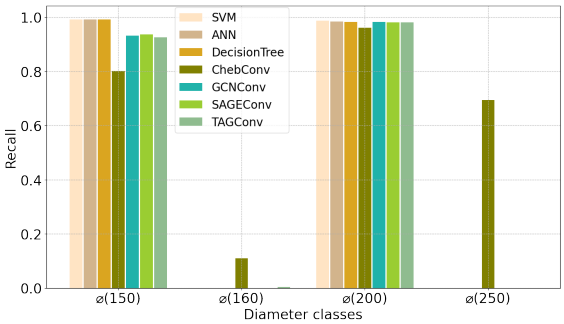
Figures 44 and 45 show the results for the Angers and Montpellier datasets, respectively. Despite having difficulties with classes with small occurrences, Strahler's order helps the models identify more classes than the dominant ones. Non-topological models SVM, Decision Tree, and MLP are unable to distinguish minor classes for the Angers dataset. Nevertheless, they predict some minor classes such as the class $\phi(400)$ with a high recall score for the Montpellier dataset (Figures 45a and 45c), despite having only 37 occurrences for this class. Unlike non-topological models, GCN models, namely, ChebConv and TAGConv, predict more classes for both datasets. Thus, GCN models outperform non-topological ones in terms of the number of detected classes.

In fact, ChebConv outperforms all models for both diameter and material prediction: this model predicted 30% of missing diameter classes $\phi(150)$ and $\phi(250)$ for the Angers dataset respectively with a recall of 79% and 77% (Figure 44c) despite having only 123 and 12 occurrences for these classes. In the case of the Montpellier dataset, ChebConv, while using only 30% of the available data, completes missing $\phi(150)$ and $\phi(300)$ diameter classes with respectively 63% and 58% recall (Figure 45a). The metric is improved when the training set is increased to 70%, thus reaching 77% and 85% respectively for these classes (Figure 45c). In comparison, the other models fail to detect these two classes for both datasets, except for TAGConv which has a very low score for the class $\phi(150)$ (Figures 44a, 44c, 45a and 45c).

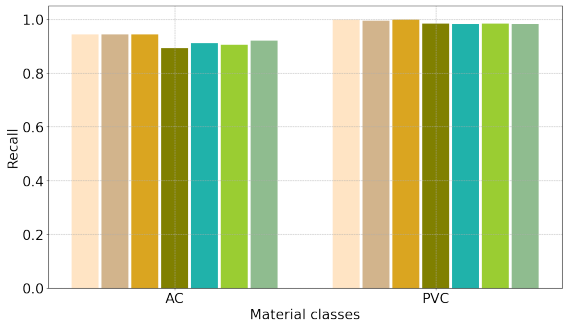
Similar results are obtained for material prediction. Indeed, besides having higher scores for both datasets, only GCN models predicted the AC class for Angers (Figures 44b and 44d). This shows that the structure of the graph and the choice of the GCN model have a great impact on the learning process.

3.3.3.2 Configuration 2

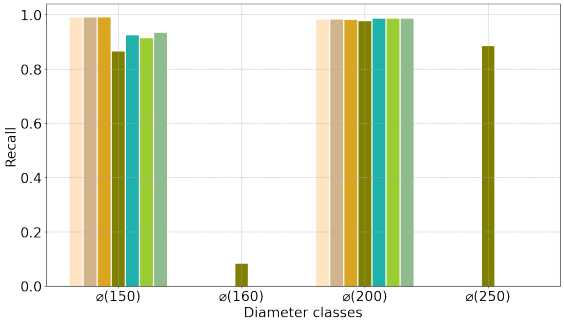
In addition to the information used in the previous configuration, the attribute "material type" is added to help predict the attribute "diameter" and vice versa. The correlation between these attributes is 0.74 for the subgraph of Angers and 0.43 for the subgraph of Montpellier. Adding this information to the models substantially increases their performance regarding the number of detected classes and the recall scores. First, except for ChebConv that already identified all the classes in the previous configuration, the number of predicted classes increases for all models. For instance, the non-topological models predict the AC class for the Angers dataset (Figure 46b). Second, Figures 46 and 47, show that recall scores have increased for the majority of the classes using the various models. Still, ChebConv outperforms all models by predicting missing values with high scores for almost all classes including the minor ones: using only 30% of the available data it achieved 80% for the class



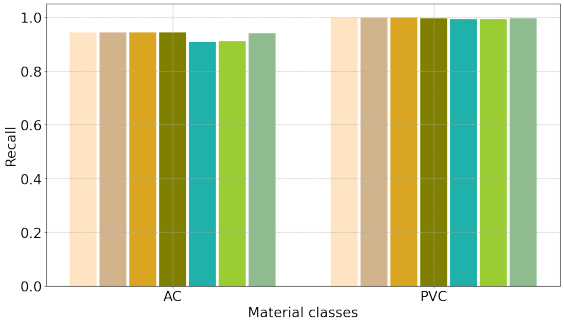
(a) Diameter prediction: 30% training and 70% test.



(b) Material prediction: 30% training and 70% test.



(c) Diameter prediction: 70% training and 30% test.

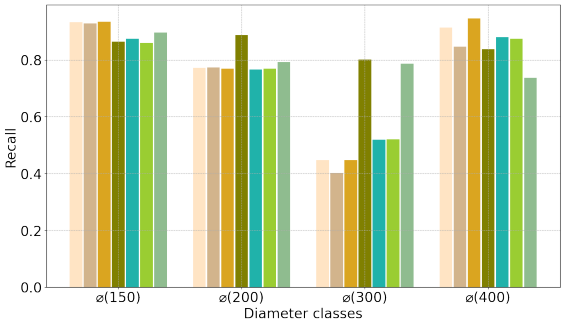


(d) Material prediction: 70% training and 30% test.

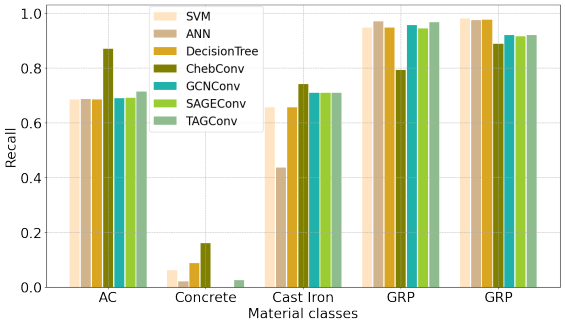
Figure 46 Configuration 2: Diameter and Material prediction for the Angers dataset for each class of the two attributes, evaluated using Recall score.

$\phi(300)$, having 34 occurrences (Figure 47a) and 70% for the class $\phi(250)$, having only 12 occurrences (Figure 46a).

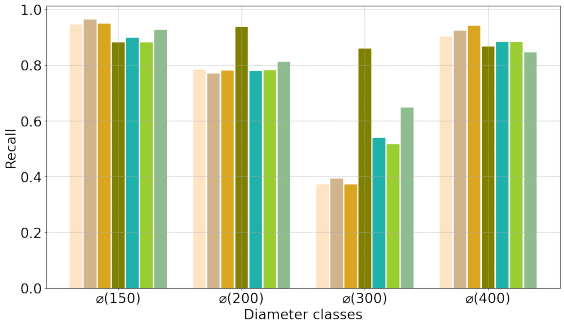
Tables 4 and 5 display the scores, Macro-Recall (MR), Macro-Precision (MP), and Macro-F1 Score (MF1) for each attribute of the two datasets of Angers and Montpellier for configurations 1 and 2 respectively, and the nine different percentages of the dataset used for training. First, for configuration 1, for both cities, Tables 4a and 4b show, as indicated before, a poor performance of the non-topological models. This was to be expected since they use only Strahler's order to distinguish the different classes, while graph models use the adjacency matrix. As for configuration 2, the scores increase for all models. Thus, the performance of non-topological models relies only on the correlations (Table 6) between Strahler's order and the targeted attributes. Second, except for ChebConv, whose performance increases when the portion of missing values decreases, all the models' performances are generally constant in configuration 1 for the Angers dataset (Table 4a) since they predict only the dominant classes. This is also to be expected for non-topological models, since there is no correlation between Strahler's order and both attributes, diameter, and material, for this dataset. However, for the Montpellier dataset, (Table 4b) where the correlation between material and Strahler is 0.08 and between diameter and Strahler



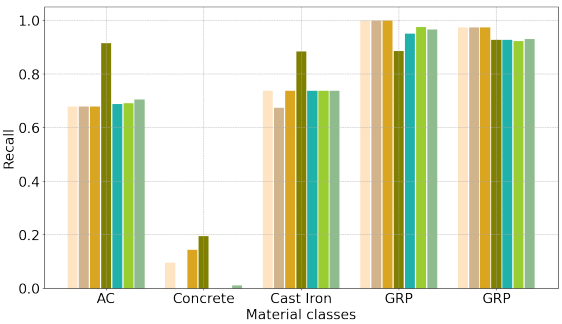
(a) Diameter prediction: 30% training and 70% test.



(b) Material prediction: 30% training and 70% test.

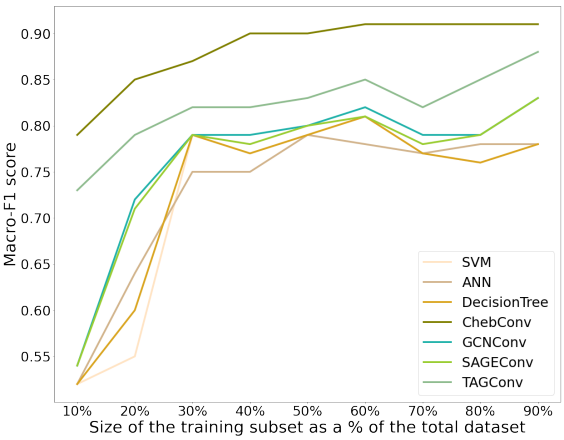


(c) Diameter prediction: 70% training and 30% test.

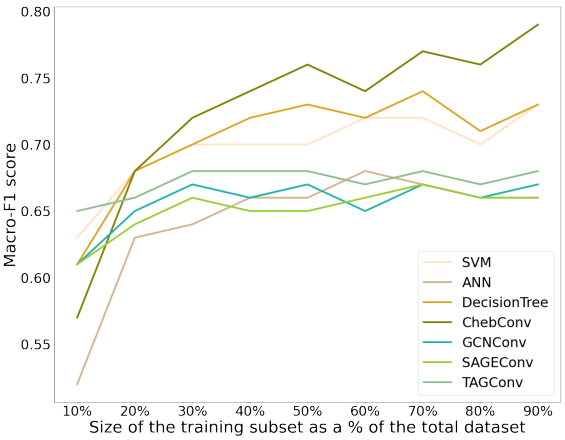


(d) Material prediction: 70% training and 30% test.

Figure 47 Configuration 2: Diameter and Material prediction for the Montpellier dataset for each class of the two attributes, evaluated using Recall score.



(a) Attribute Diameter.



(b) Attribute Material.

Figure 48 Models performances evolution (F1 score) while decreasing the amount of missing data for the configuration 2 of the Montpellier dataset.

Table 4 **Configuration 1.** Results obtained for Angers and Montpellier dataset by the seven models in terms of Macro-Recall (MR), Macro-Precision (MP) and Macro-F1 (MF1) scores, for the two classes and with different percentages of the dataset used for training.

(a) Angers dataset

Angers																						
		SVM			ANN			DT			ChebConv			GCNConv			SAGEConv			TAGConv		
Attribute	%	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1
Diameter	10	0.25	0.19	0.22	0.25	0.19	0.21	0.25	0.19	0.22	0.41	0.6	0.45	0.26	0.28	0.23	0.25	0.21	0.22	0.27	0.34	0.26
	20	0.25	0.19	0.22	0.25	0.19	0.22	0.25	0.19	0.22	0.51	0.69	0.56	0.26	0.28	0.24	0.26	0.24	0.23	0.29	0.38	0.29
	30	0.25	0.19	0.22	0.25	0.19	0.22	0.25	0.19	0.22	0.57	0.77	0.63	0.26	0.26	0.23	0.26	0.27	0.23	0.3	0.4	0.3
	40	0.25	0.19	0.22	0.25	0.19	0.22	0.25	0.19	0.22	0.61	0.79	0.66	0.26	0.26	0.23	0.26	0.26	0.23	0.3	0.42	0.3
	50	0.25	0.19	0.22	0.25	0.19	0.22	0.25	0.19	0.22	0.66	0.8	0.7	0.26	0.28	0.24	0.27	0.31	0.25	0.3	0.41	0.3
	60	0.25	0.19	0.22	0.25	0.19	0.22	0.25	0.19	0.22	0.68	0.81	0.71	0.25	0.23	0.23	0.27	0.36	0.26	0.3	0.41	0.31
	70	0.25	0.19	0.22	0.25	0.19	0.22	0.25	0.19	0.22	0.69	0.77	0.71	0.26	0.29	0.23	0.27	0.35	0.25	0.3	0.4	0.3
	80	0.25	0.19	0.22	0.25	0.19	0.22	0.25	0.19	0.22	0.69	0.78	0.72	0.26	0.28	0.23	0.26	0.29	0.24	0.3	0.41	0.3
	90	0.25	0.19	0.22	0.25	0.19	0.22	0.25	0.19	0.22	0.75	0.76	0.74	0.27	0.29	0.24	0.26	0.27	0.23	0.31	0.43	0.31
Material	10	0.5	0.42	0.45	0.5	0.41	0.45	0.49	0.43	0.45	0.62	0.78	0.66	0.54	0.69	0.53	0.52	0.63	0.5	0.56	0.73	0.57
	20	0.5	0.41	0.45	0.5	0.41	0.45	0.5	0.41	0.45	0.71	0.82	0.75	0.53	0.66	0.51	0.53	0.6	0.5	0.59	0.83	0.61
	30	0.5	0.41	0.45	0.5	0.41	0.45	0.5	0.42	0.45	0.78	0.86	0.81	0.54	0.74	0.53	0.54	0.71	0.53	0.59	0.81	0.6
	40	0.5	0.41	0.45	0.5	0.41	0.45	0.5	0.41	0.45	0.85	0.89	0.86	0.54	0.76	0.53	0.54	0.74	0.53	0.6	0.82	0.63
	50	0.5	0.41	0.45	0.5	0.41	0.45	0.5	0.41	0.45	0.86	0.88	0.87	0.54	0.76	0.53	0.55	0.77	0.54	0.6	0.87	0.63
	60	0.5	0.41	0.45	0.5	0.41	0.45	0.5	0.41	0.45	0.87	0.9	0.89	0.53	0.64	0.5	0.54	0.73	0.53	0.6	0.83	0.63
	70	0.5	0.41	0.45	0.5	0.41	0.45	0.5	0.41	0.45	0.87	0.92	0.89	0.52	0.6	0.49	0.52	0.61	0.49	0.61	0.85	0.63
	80	0.5	0.42	0.45	0.5	0.42	0.45	0.5	0.42	0.45	0.88	0.93	0.9	0.53	0.75	0.52	0.53	0.75	0.52	0.6	0.87	0.63
	90	0.5	0.41	0.45	0.5	0.41	0.45	0.5	0.41	0.45	0.91	0.95	0.93	0.53	0.64	0.5	0.53	0.64	0.5	0.62	0.91	0.65

(b) Montpellier dataset

Montpellier																						
		SVM			ANN			DT			ChebConv			GCNConv			SAGEConv			TAGConv		
Attribute	%	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1
Diameter	10	0.22	0.16	0.18	0.23	0.16	0.19	0.22	0.17	0.19	0.48	0.62	0.52	0.23	0.38	0.23	0.23	0.37	0.23	0.38	0.52	0.41
	20	0.25	0.17	0.2	0.36	0.24	0.29	0.38	0.26	0.31	0.63	0.74	0.67	0.39	0.44	0.37	0.39	0.47	0.37	0.43	0.5	0.43
	30	0.48	0.31	0.38	0.48	0.31	0.38	0.48	0.31	0.38	0.7	0.78	0.73	0.47	0.39	0.39	0.47	0.43	0.4	0.44	0.47	0.42
	40	0.48	0.32	0.38	0.48	0.32	0.38	0.48	0.32	0.38	0.76	0.85	0.79	0.46	0.37	0.39	0.46	0.37	0.39	0.44	0.52	0.42
	50	0.48	0.31	0.38	0.48	0.31	0.38	0.48	0.31	0.38	0.8	0.88	0.83	0.47	0.34	0.39	0.47	0.35	0.39	0.47	0.53	0.43
	60	0.48	0.32	0.38	0.48	0.32	0.38	0.48	0.32	0.38	0.83	0.88	0.84	0.48	0.33	0.39	0.48	0.35	0.39	0.46	0.53	0.42
	70	0.47	0.32	0.38	0.47	0.32	0.38	0.47	0.32	0.38	0.85	0.91	0.87	0.47	0.34	0.38	0.47	0.34	0.38	0.45	0.46	0.41
	80	0.49	0.34	0.4	0.49	0.34	0.4	0.49	0.34	0.4	0.85	0.91	0.87	0.48	0.34	0.4	0.48	0.35	0.4	0.48	0.55	0.44
	90	0.47	0.33	0.38	0.47	0.33	0.38	0.47	0.33	0.38	0.87	0.91	0.88	0.47	0.33	0.38	0.47	0.33	0.38	0.46	0.5	0.41
Material	10	0.36	0.33	0.33	0.35	0.29	0.31	0.36	0.33	0.32	0.43	0.55	0.45	0.36	0.33	0.34	0.36	0.33	0.33	0.35	0.36	0.34
	20	0.39	0.33	0.34	0.39	0.31	0.33	0.39	0.32	0.34	0.55	0.68	0.57	0.39	0.35	0.35	0.4	0.36	0.36	0.39	0.37	0.36
	30	0.39	0.32	0.35	0.39	0.28	0.32	0.39	0.33	0.35	0.6	0.69	0.62	0.4	0.37	0.36	0.4	0.35	0.36	0.41	0.39	0.37
	40	0.39	0.33	0.35	0.39	0.31	0.33	0.39	0.33	0.35	0.64	0.75	0.65	0.39	0.33	0.35	0.39	0.34	0.35	0.41	0.38	0.38
	50	0.39	0.32	0.34	0.39	0.27	0.31	0.39	0.33	0.34	0.68	0.84	0.71	0.39	0.33	0.35	0.39	0.33	0.34	0.42	0.4	0.38
	60	0.39	0.33	0.34	0.39	0.3	0.32	0.39	0.33	0.34	0.72	0.85	0.76	0.39	0.33	0.35	0.39	0.34	0.35	0.42	0.38	0.38
	70	0.39	0.32	0.34	0.39	0.3	0.33	0.39	0.32	0.34	0.72	0.83	0.75	0.39	0.33	0.35	0.4	0.35	0.35	0.42	0.36	0.38
	80	0.38	0.33	0.33	0.38	0.29	0.32	0.38	0.33	0.33	0.72	0.88	0.75	0.37	0.31	0.33	0.38	0.32	0.34	0.41	0.36	0.37
	90	0.39	0.33	0.33	0.38	0.28	0.31	0.39	0.34	0.32	0.74	0.78	0.75	0.4	0.35	0.36	0.39	0.33	0.34	0.44	0.4	0.41

is 0.31, the non-topological models' performances increase when the percentage of missing data decreases for the attribute diameter. Also, in configuration 2, the models' performances evolve differently for the two datasets. For Angers, all models are nearly constant, although a small increase can be noted in ChebConv's performance while the missing data decreases. These scores (Table 5a) can be explained by the high correlation of the attributes material and diameter (0.74). For the Montpellier dataset, where the correlation is lower compared to the Angers dataset, almost all the models' performances increase. Figure 48 illustrates this evolution using the Macro-F1Score metric. The differences in performance related to the GCN models are detailed in the next paragraph.

Our experiments show that for real-world configurations, ChebConv yields the best results for both datasets and both predicted attributes. Spatial approaches fail to distinguish minority classes compared to the spectral approaches (i.e., ChebConv) and

Table 5 **Configuration 2**. Results obtained for the Angers and Montpellier datasets by the seven models in terms of Macro-Recall (MR), Macro-Precision (MP) and Macro-F1 (MF1) scores, for the two classes and with different percentages of the dataset used for training.

(a) Angers dataset

Angers																						
		SVM			ANN			DT			ChebConv			GCNConv			SAGEConv			TAGConv		
Attribute	%	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1
Diameter	10	0.49	0.46	0.47	0.49	0.46	0.48	0.49	0.46	0.48	0.58	0.76	0.62	0.48	0.45	0.46	0.47	0.45	0.46	0.48	0.5	0.47
	20	0.49	0.46	0.48	0.49	0.46	0.48	0.49	0.46	0.48	0.64	0.74	0.66	0.48	0.45	0.46	0.48	0.45	0.46	0.48	0.46	0.47
	30	0.49	0.46	0.48	0.49	0.46	0.48	0.49	0.46	0.48	0.64	0.73	0.66	0.48	0.45	0.47	0.48	0.45	0.47	0.48	0.46	0.47
	40	0.49	0.46	0.48	0.49	0.46	0.47	0.49	0.46	0.47	0.68	0.76	0.7	0.48	0.45	0.47	0.48	0.45	0.47	0.49	0.46	0.47
	50	0.49	0.46	0.48	0.49	0.46	0.48	0.49	0.46	0.48	0.67	0.79	0.69	0.48	0.45	0.47	0.48	0.45	0.47	0.48	0.46	0.47
	60	0.49	0.46	0.48	0.49	0.46	0.48	0.49	0.46	0.48	0.68	0.8	0.7	0.48	0.45	0.47	0.48	0.45	0.47	0.48	0.46	0.47
	70	0.49	0.46	0.48	0.49	0.46	0.47	0.49	0.46	0.48	0.7	0.77	0.71	0.48	0.46	0.47	0.47	0.46	0.47	0.48	0.46	0.47
	80	0.5	0.46	0.48	0.5	0.46	0.48	0.5	0.46	0.48	0.66	0.71	0.66	0.48	0.45	0.46	0.48	0.45	0.46	0.48	0.46	0.47
	90	0.49	0.46	0.48	0.49	0.46	0.48	0.49	0.46	0.48	0.74	0.77	0.74	0.48	0.45	0.47	0.48	0.46	0.47	0.48	0.46	0.47
Material	10	0.96	0.99	0.97	0.96	0.99	0.97	0.97	0.99	0.98	0.87	0.93	0.9	0.93	0.95	0.94	0.92	0.95	0.94	0.9	0.94	0.92
	20	0.96	0.99	0.98	0.97	0.99	0.98	0.97	0.99	0.98	0.91	0.94	0.92	0.94	0.96	0.95	0.94	0.95	0.95	0.94	0.96	0.95
	30	0.97	0.99	0.98	0.97	0.99	0.98	0.97	0.99	0.98	0.94	0.96	0.95	0.95	0.95	0.95	0.94	0.96	0.95	0.95	0.96	0.96
	40	0.97	0.99	0.98	0.97	0.99	0.98	0.97	0.99	0.98	0.94	0.96	0.95	0.94	0.96	0.95	0.94	0.95	0.95	0.95	0.96	0.96
	50	0.97	0.99	0.98	0.97	0.99	0.98	0.97	0.99	0.98	0.95	0.97	0.96	0.95	0.96	0.95	0.94	0.96	0.95	0.96	0.98	0.97
	60	0.98	0.99	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.97	0.97	0.97	0.96	0.97	0.96	0.96	0.97	0.96	0.97	0.98	0.97
	70	0.97	0.99	0.98	0.97	0.99	0.98	0.97	0.99	0.98	0.97	0.99	0.98	0.95	0.97	0.96	0.95	0.97	0.96	0.97	0.99	0.98
	80	0.96	0.99	0.98	0.96	0.99	0.98	0.96	0.99	0.98	0.95	0.98	0.96	0.94	0.96	0.95	0.94	0.96	0.95	0.96	0.98	0.97
	90	0.95	0.99	0.97	0.95	0.99	0.97	0.95	0.99	0.97	0.97	0.97	0.97	0.93	0.96	0.94	0.93	0.96	0.94	0.95	0.99	0.97

(b) Montpellier dataset

Montpellier																						
		SVM			ANN			DT			ChebConv			GCNConv			SAGEConv			TAGConv		
Attribute	%	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1	MR	MP	MF1
Diameter	10	0.5	0.64	0.52	0.51	0.58	0.52	0.51	0.6	0.52	0.75	0.86	0.79	0.52	0.64	0.54	0.52	0.61	0.54	0.67	0.83	0.73
	20	0.54	0.64	0.55	0.64	0.69	0.64	0.59	0.68	0.6	0.82	0.89	0.85	0.7	0.81	0.72	0.68	0.82	0.71	0.77	0.84	0.79
	30	0.77	0.86	0.79	0.74	0.81	0.75	0.77	0.85	0.79	0.85	0.9	0.87	0.76	0.85	0.79	0.76	0.84	0.79	0.8	0.86	0.82
	40	0.77	0.85	0.77	0.77	0.81	0.75	0.78	0.84	0.77	0.88	0.92	0.9	0.79	0.83	0.79	0.78	0.82	0.78	0.81	0.85	0.82
	50	0.77	0.86	0.79	0.78	0.85	0.79	0.79	0.85	0.79	0.88	0.92	0.9	0.79	0.83	0.8	0.79	0.84	0.8	0.81	0.88	0.83
	60	0.8	0.86	0.81	0.77	0.83	0.78	0.8	0.85	0.81	0.9	0.93	0.91	0.82	0.83	0.82	0.82	0.83	0.81	0.85	0.86	0.85
	70	0.75	0.85	0.77	0.76	0.84	0.77	0.76	0.85	0.77	0.89	0.93	0.91	0.77	0.82	0.79	0.77	0.82	0.78	0.81	0.86	0.82
	80	0.75	0.81	0.76	0.77	0.86	0.78	0.75	0.81	0.76	0.89	0.93	0.91	0.79	0.82	0.79	0.79	0.82	0.79	0.85	0.87	0.85
	90	0.79	0.81	0.78	0.79	0.8	0.78	0.79	0.8	0.78	0.89	0.96	0.91	0.84	0.85	0.83	0.84	0.83	0.83	0.9	0.87	0.88
Material	10	0.6	0.72	0.63	0.54	0.52	0.52	0.6	0.68	0.61	0.54	0.73	0.57	0.6	0.65	0.61	0.6	0.64	0.61	0.63	0.7	0.65
	20	0.66	0.73	0.68	0.63	0.65	0.63	0.66	0.76	0.68	0.65	0.78	0.68	0.64	0.68	0.65	0.63	0.68	0.64	0.65	0.71	0.66
	30	0.67	0.79	0.7	0.62	0.71	0.64	0.67	0.81	0.7	0.69	0.81	0.72	0.65	0.7	0.67	0.65	0.69	0.66	0.67	0.73	0.68
	40	0.67	0.77	0.7	0.64	0.7	0.66	0.68	0.82	0.72	0.71	0.82	0.74	0.65	0.69	0.66	0.64	0.68	0.65	0.66	0.72	0.68
	50	0.68	0.76	0.7	0.65	0.7	0.66	0.7	0.83	0.73	0.73	0.85	0.76	0.67	0.7	0.67	0.65	0.67	0.65	0.67	0.71	0.68
	60	0.69	0.81	0.72	0.67	0.72	0.68	0.69	0.84	0.72	0.72	0.81	0.74	0.64	0.69	0.65	0.65	0.69	0.66	0.66	0.71	0.67
	70	0.7	0.83	0.72	0.66	0.7	0.67	0.71	0.87	0.74	0.76	0.81	0.77	0.66	0.7	0.67	0.66	0.69	0.67	0.67	0.72	0.68
	80	0.67	0.78	0.7	0.65	0.71	0.66	0.68	0.82	0.71	0.73	0.84	0.76	0.65	0.69	0.66	0.65	0.69	0.66	0.66	0.72	0.67
	90	0.72	0.78	0.73	0.65	0.69	0.66	0.72	0.78	0.73	0.79	0.8	0.79	0.68	0.69	0.67	0.68	0.67	0.66	0.68	0.7	0.68

slightly outperform non-topological approaches. The fact that SAGEConv, which is a spatial approach, has a nearly similar evolution performance as non-topological models, and is outperformed by ChebConv, may be explained by the fixed-size set of the neighbourhood, where not all the neighbourhoods are explored. Furthermore, for the spectral approaches, ChebConv surpassing GCNConv may be explained by the differences in the number of K-localised filters since GCNConv uses only $K = 1$ to avoid overfitting. To confirm this assumption we varied the values of parameter K to 1,10,15 and 20 for the ChebConv model and compared the new experiments to the GCNConv. Figure 49 shows that GCNConv and ChebConv with $K = 1$ have similar performances regarding the number of predicted classes and the recall scores, when predicting the diameter values for the Montpellier dataset. Moreover, comparing the performance of ChebConv with different K values shows that increasing the number

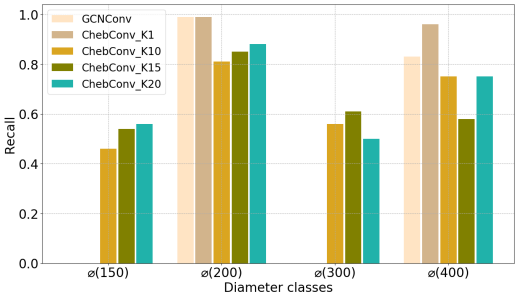
Table 6 Attributes correlations.

Angers dataset

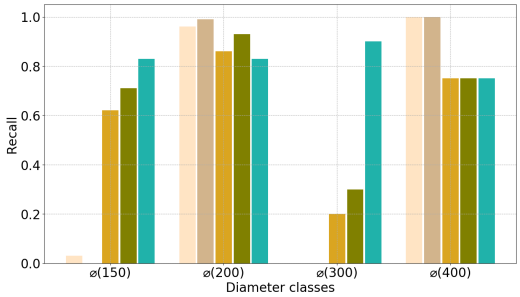
Attributes	Diameter	Material	Strahler
Diameter	1	0.74	0.06
Material	0.74	1	0.01
Strahler	0.06	0.01	1

Montpellier dataset

Attributes	Diameter	Material	Strahler
Diameter	1	0.43	0.31
Material	0.43	1	0.08
Strahler	0.31	0.08	1



(a) Diameter prediction: 30% training and 70% test.



(b) Diameter prediction: 70% training and 30% test.

Figure 49 Comparing GCNConv model with ChebConv model on the Montpellier dataset.

of neighbour nodes used in the learning process improves the prediction results. This was also noted for TAGConv.

3.3.4 Discussion and conclusions

This study was conducted to investigate whether machine learning algorithms can be used for Missing Value Imputation on wastewater networks. We carried out tests using 7 different models; four Graph Convolutional Network models: GCN, ChebNet, TAGCN, and GraphSAGE, and three popular non-topological models: SVM, Decision Trees, and a MultiLayer Perceptron. The results show that machine learning models are an efficient tool for completing missing attributes for wastewater networks when various types of information about a network are available. This is highlighted in the second test configuration we explored. Moreover, for extreme situations, when only the network layout and partial attribute information are available (i.e., the first test configuration), the ChebConv spectral GCN approach, which is based on the approximation of the spectrum of the graph Laplacian, yields the best results for the completion of attribute values in general, and minority classes in particular. ChebConv yields acceptable results also when a small percentage of the available data is used for training. This was demonstrated in several studies using GCN-based models. (Spinelli et al., 2020) showed that, in comparison with other approaches such as KNN, the performance of their GCN-based model increased substantially when the percentage of the missing data increased. In a different application, similar con-

clusions were reached by (Rahimi et al., 2018) when inferring users' geo-localisation in social media. The authors used a semi-supervised configuration combining graph structure and text and showed that a GCN-based model performs well under minimal supervision scenarios by effectively using unlabelled data.

The machine learning models that we used in this application require specific conditions. First, the classes to be learnt must be part of the training dataset. We complied with this request by ignoring classes with less than 10 occurrences. However, this led to fewer minority classes in the test subset and therefore impacted the prediction results substantially. Second, machine learning models are known to require important data quantity to achieve satisfying results. Having achieved these scores while using such restricted datasets shows that this approach can be even more promising with larger datasets. We would like also to emphasize that our objective was not to determine the best GCN architecture for wastewater network data completion, but rather to investigate the impact of the structure of the graph as a learning factor on the prediction results. In this study, we used the default implementation of the GCN models as described in the original papers. Although these models showed excellent performance in various domains such as information science, bibliometrics, water distribution systems, or biology (Du et al., 2017; Hamilton et al., 2017; Kipf & Welling, 2017; Tsiami & Makropoulos, 2021), they can be further adapted to the specific context of each domain to produce better results. For instance, in (Jepsen et al., 2019), a novel type of GCN for road networks called Relational Fusion Network (RFN) is put forward for driving speed estimation and speed limit classification. The results indicate that RFN outperforms state-of-the-art GCN algorithms such as GraphSAGE in this application.

To assess whether the structure of the graph, modelled in our case by the adjacency matrix, has an impact on the learning process, non-topological models were trained using only the available attributes. That is Strahler's order for the first configuration and Strahler's order, diameter, and material for the second configuration. Strahler's order is used as a proxy for network topology in these models. For the GCN models, in addition to these attributes, the adjacency matrix is required and is also provided. The matrix is not used for the non-topological models because they are not built to deal with graph structures and require a pre-processing step to operate. This consists in representing or encoding the graph in a suitable form for the targeted model. As stated in Section 3.3.1, this operation is complex and does not guaranty the full use of the graph structure, while GCN models can easily handle information such as adjacency or angle between pipes to perform MVI operations. Therefore, no pre-processing was carried out in this work.

The attributes diameter, material, and Strahler's order were used only as illustration examples in this study. We aim to show that machine learning models can be an efficient method to help all entities facing the problem of missing wastewater network data, to overcome this challenge. The use of both numerical (diameter) and categorical (material) attributes shows that this approach overcomes the limits of the statistical methods used in (Kabir et al., 2020). In some instances, Strahler's order, which is dependent on the dataset, may not be the best descriptor. For instance, since the Angers dataset is very small, the pipe diameters do not increase when moving from the upstream wastewater catchments to the vicinity of the treatment plant. This leads to a lack of correlation between Strahler's order and diameter (Table 6). Thus, Strahler's order does not affect the diameter predictions for the Angers dataset, contrary to the Montpellier network. One may also use the type of buildings near the pipes as an attribute to predict their diameter. The main idea is that, since network construction rules vary from one country to another, and between regions of the same country, machine learning models can easily integrate new information to make predictions and improve them. It all depends on the available data and knowledge about the targeted network.

Urban managers and environmental monitoring services are often faced with incomplete datasets and have to resort to Missing Value Imputation (MVI) or Missing Data Imputation (MDI) algorithms. GCN models would provide managers with an additional accessible resource to overcome data imperfection challenges and support decision-makers, be it to conduct repairs, predict future damages such as in (Kumar et al., 2018) or run a hydraulic simulation model. Indeed, several urban utility networks such as gas, water, and electrical supplies are structured as graphs with nodes and edges. Our proposition would help asset management tasks by providing a better estimate for given characteristics of the undocumented portions of the network. Another important feature of Smart City management plans is air and water pollution monitoring. Given the spatial and temporal variability of environmental indicators, these monitoring plans rely on a network of sensors, spread out over large geographical areas. As with any piece of equipment, these devices are prone to failure and damage, resulting in missing data. By resorting to GNNs, managers would be able to get the most of their network's structure and get more accurate estimations of the missing data. They would thus be able to better inform citizens and improve their quality of life.

3.4. Graph Neural Networks for pipe prediction

3.4.1 Context

This work is dedicated to the problem of missing spatial objects of wastewater network databases. Manhole covers are visible on the ground, their positions can either be collected manually by the operators or be identified automatically, for instance on high resolution (Commandre et al., 2017) or Google Street View images (Boller et al., 2019). However, pipes being buried make their inspection challenging. As stated previously, operators often resort to signal/radar based methods to extract information about the buried objects, which is a complex task that is not always affordable. As an alternative, one may use the manhole positions to map the networks automatically. To the best of our knowledge, this problem was addressed previously only in (Chahinian et al., 2019), where an optimisation algorithm was put forward to automatically create possible wastewater network maps using manhole cover positions. In the next paragraph we summarize and discuss the limits of the work in (Chahinian et al., 2019).

An optimisation method for mapping wastewater networks (Chahinian et al., 2019). The main steps of this proposition can be summarized in the following:

1. Given a set $S = \{M_1, \ldots, M_N\}$ of N manhole cover positions, a subset of possible connections are created using a Delaunay triangulation. In Euclidian distance, a Delaunay triangulation can be proven to be a spanning graph of set S. However, this rule has no set rationale for wastewater networks. Thus, additional possible connections located within a radius set by the user are added.

2. A cost value $c(M_i, M_j)$ is assigned to each possible connection according to the following cost function:

$$c(M_i, M_j) = \alpha_L C_{L_{i,j}} + \alpha_S C_{S_{i,j}} + \alpha_\theta C_{\theta_{i,j}} + P_r + P_b \tag{3.9}$$

   where $C_{L_{i,j}}, C_{S_{i,j}}$ and $C_{\theta_{i,j}}$ are the costs associated respectively to the length, the slope and the angle of a given connection between M_i and M_j. α_L, α_S and α_θ are the weights associated respectively to the length, the slope and the angle costs. These costs are defined based on domain knowledge. For example in France, the maximal distance recommended between two covers is 80 m, thus for C_L, it is assumed that if the length $L_{i,j}$ of a connection is greater than 160 m, $C_{L_{i,j}} = 1$, otherwise $C_{L_{i,j}} = L_{i,j}/160$. P_r

is defined as the length of edge that is outside a road divided by a distance $d = 20$ m, since the pipes are often buried under the pavements or the road network. P_b is defined as the percentage of an edge that crosses a building multiplied by a number N. $N = 4$, so that P_b is greater than one if more than 25% of the edge is crossing a building. This is based on the hypothesis that public pipes are usually not buried under buildings.

3. Starting from the node that represents the outfall, an optimisation algorithm was put forward to choose the connections with the lowest costs, and to ensure gravity fed flow based on the slope information extracted from a Digital Elevation Model (DEM).

This method was tested using data from two small towns in the south of France: Prades-Le-Lez and Ramonville-Saint-Agne. Although the results were encouraging this method suffers from two main issues:

- Since the pipes are chosen iteratively starting from the outfall, errors may propagate to the entire network.
- This method is based on the assumption of gravity fed flow. However, slope is the most difficult attribute to acquire and often unavailable in public databases. It is thus derived from elevation values, which are not always accurate and may lead to wrong flow directions.

Given the results obtained in the previous contribution (Section 3.3). Here, we investigate whether GNNs can help map the networks. We propose a new GNN model dedicated to predicting the positions of the pipes automatically. First, artificial neural networks predict instances independently, hence errors propagation is not an issue. Second, slope attribute will not be used as part of the input properties of the GNN model. In fact, wastewater networks can often be divided into multiple sub-networks, each having their independent outfall. Usually these sub-networks are installed following the same rules and regulations especially when they are built on the same period, such as identical average distance between manholes. Moreover, even for poorly documented networks, some data are often available, especially in medium and large cities. In this study, we aim to investigate whether we can learn from the available data to predict pipe presence between the manholes. Our proposition is described in the following section.

3.4.2 Materials and methods

3.4.2.1 Wastewater Graph Neural Network

We aim to learn to distinguish between negative and positive connections, where the positives indicate real pipes between the manholes. Our proposition is inspired by GraphSAGE (Hamilton et al., 2017). We chose the GrapheSAGE framework due to its ability to learn not only from fixed graphs but to generalize to unseen nodes. In fact, contrary to popular approaches such as GCN (Kipf & Welling, 2017) or ChebNet (Defferrard et al., 2017) where the focus is on learning embedding vectors for each node using the neighbors' information, GraphSAGE learns an aggregation function from the nodes' neighborhood. This allows to use the learned aggregation to generate the embedding of unseen nodes at the test step.

Let $G = (V, E)$ be an undirected graph that represents the graph of the possible connections, where $V = \{v_1, \ldots, v_n\}$ is the set of n nodes that represent the n potential pipes. E is the set of edges, where the edge $e_{i,j} = (v_i, v_j) \in E$ is a couple of pipes having one common node. The neighborhood of a node v_i is defined as $\mathcal{N}(v_i) = \{v_j \in V, (v_i, v_j) \in E\}$. Let $X\{x_1, \ldots, x_n\} \in \mathbb{R}^{nm}$, be the set of m attributes associated to each of the n nodes, and $L = \{l_1, \ldots, l_n\} \in \{0, 1\}^n$, the set of nodes' labels, where $l_i = 1$ when the node represents a real pipe, and $l_i = 0$ otherwise.

We define the message received by each node v_i at iteration $k \in K$, where K is the number of iterations, as follows:

$$h_{v_i}^k = W_1 h_{v_i}^{k-1} + W_2 \underset{v_j \in \mathcal{N}(v_i)}{\text{Mean}} \left(w_{i,j} h_{v_j}^{k-1} \right) \tag{3.10}$$

where $W1$ and $W2$ are sets of learnable parameters. $h_{v_i}^k$ is the embedding of the node v_i at iteration k, with $h_{v_i}^0 = x_i, \forall v_i \in V$; $w_{i,j}$ is the edge attribute computed from the angle between nodes v_i and v_j as explained in the following paragraph. Algorithm 2, describes the embedding learning steps.

Algorithm 2 Wastewater Graph Neural Network

$h_{v_i}^0 = x_i, \forall v_i \in V$;
for $k = 1 \ldots K$ **do**
    for $v_i \in V$ **do**
        $h_{v_i}^k = W_1 h_{v_i}^{k-1} + W_2 \underset{v_j \in \mathcal{N}(v_i)}{\text{Mean}} w_{i,j} h_{v_j}^{k-1}$
    end for
end for

The weight $w_{i,j}$ between the edges is the most important parameter in our proposition, since its definition is based on prior knowledge related to the construction of the

wastewater network. The role of the weights is to promote certain relationships more than others. We defined three different weights for three different layers:

Weight1: as outlined in chapter 2, wastewater networks are organized in the form of strokes, where a stroke describes a "good continuity" between lines that do not exceed a degree of deviation. Thus, it is likely that when a node represents a real pipe, the neighbors with which it forms an angle close to 180 degrees represent also real pipes. To promote this relationship we define the first transformation function (Figure 50a) as follows:

$$w1_{i,j} = \begin{cases} 0, & \text{if } \widehat{v_i v_j} \leq 160 \\ \cos((2\pi/90)\widehat{v_i v_j}), & \text{otherwise} \end{cases} \tag{3.11}$$

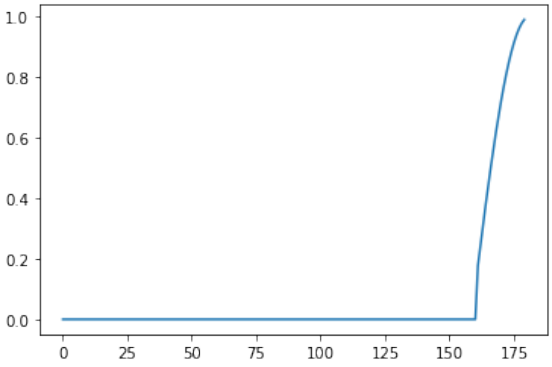
the message received by each node is thus defined as:

$$h_{v_i}^k = W_1 h_{v_i}^{k-1} + W_2 \underset{v_j \in \mathcal{N}(v_i)}{\text{Mean}} \left(w1_{i,j} h_{v_j}^{k-1} \right) \tag{3.12}$$

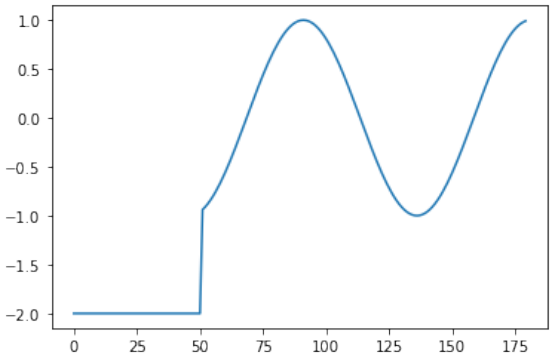
Weight2: when analyzing wastewater networks, we notice that the angles between the pipes are often close to either 90 or 180 degrees. Thus, the nodes respecting this constraint should have similar embeddings. By contrast, pipes having an angle between 0 and 50 degrees occur rarely and should not be encouraged in the learning process. In addition, the semantic of the edges in this application differs from the usual applications of GNNs. In fact, in most popular datasets such as social networks or citations, adjacent nodes are likely to belong to the same class when they have similar attributes; and adjacent nodes can all belong to the same class. However, in our application, given the way the graph of possible connections is built, when a node represents a real pipe (class =1), usually only a limited number of neighbors are also real pipes; often 2 neighbors when the network branches, 0 or 1 when there is no branching out. Thus, the desired model should separate the neighbors locally into both classes. To achieve this goal, we define a second transformation function that associates each angle to a weight as follows:

$$w2_{i,j} = \begin{cases} -2.0, & \text{if } \widehat{v_i v_j} \leq 50 \\ \cos((2\pi/90)\widehat{v_i v_j}), & \text{otherwise} \end{cases} \tag{3.13}$$

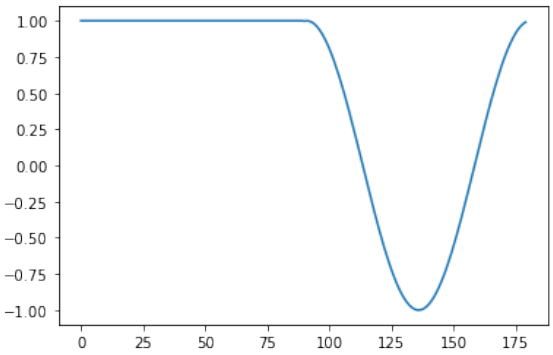
Figure 50b, shows this transformation, where the weights are close to 1 when the angles are close values of 90 and 180 degrees. The values are -2.0 when the



(a) The transformation applied as weight1.



(b) The transformation applied as weight2.



(c) The transformation applied as weight3.

Figure 50 Angle transformations into weights.

angle values are unlikely to be found in real databases. The message received by each node is then defined as:

$$h_{v_i}^k = W_1 h_{v_i}^{k-1} + W_2 \underset{v_j \in \mathcal{N}(v_i)}{\text{Mean}} \left((1 - w2_{i,j}) h_{v_j}^{k-1}\right) \tag{3.14}$$

That is, the nodes receive information about the neighbors that are unlikely to be from the same class.

Weight3: to avoid neglecting infrequent connections between the pipes, we propose a third transformation which promotes the relationships which have not been considered in the previous weights. This transformation (Figure 50c) is defined as:

$$w3_{i,j} = \begin{cases} \cos((2\pi/90)\widehat{v_i v_j}), & \text{if } 90 \leq \widehat{v_i v_j} \leq 180 \\ 1.0, & \text{otherwise} \end{cases} \tag{3.15}$$

The message received by each node is then defined as:

$$h_{v_i}^k = W_1 h_{v_i}^{k-1} + W_2 \underset{v_j \in \mathcal{N}(v_i)}{\text{Mean}} \left(w3_{i,j} h_{v_j}^{k-1}\right) \tag{3.16}$$

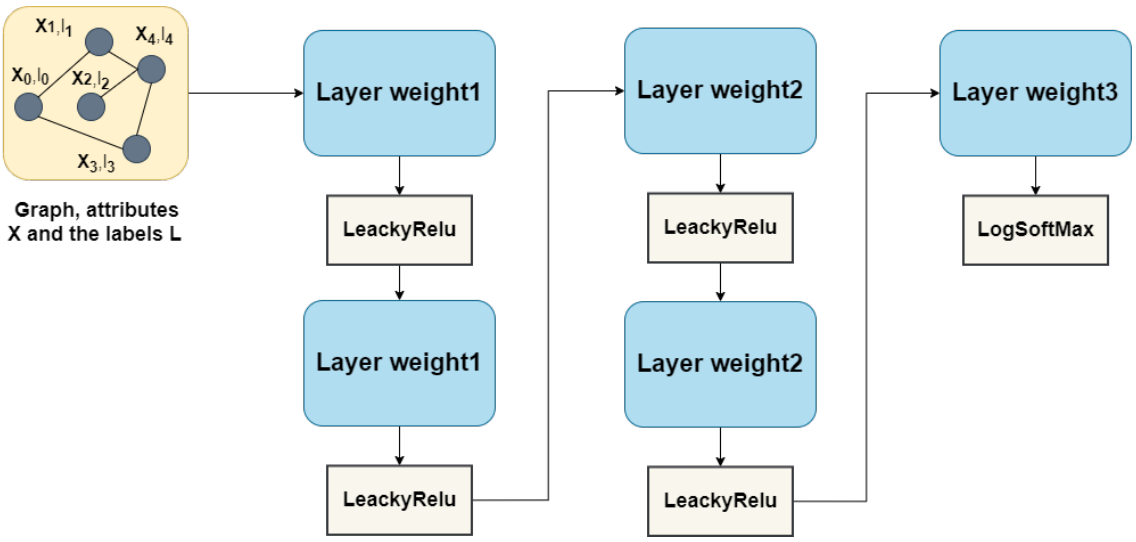


Figure 51 WaGNN: The Graph Convolutional Network architecture for pipe prediction.

We implemented our proposition using Pytorch Geometric (Fey & Lenssen, 2019). The GNN network that we use for the supervised training after tuning the hyperparameters is illustrated in Figure 51. We call this network Wastewater Graph Neural Network (WaGNN). We applied LeakyRelu as an activation function to avoid dying negative values, and logSoftmax as the activation function of the output layer. The learnable parameters W1 and W2 are modeled by Linear layers as defined in Pytorch (Paszke et al., 2017). The gradient descent is achieved using the Adam optimization algorithm with a learning rate of value 0.01, and the Cross Entropy as the loss function.

3.4.2.2 Datasets and experiment

We applied our GNN on real datasets of wastewater networks: Prades-Le-Lez and Ramonville-Saint-Agne, two towns located in the south of France. In fact, we adapted the same graphs of possible connections used in (Chahinian et al., 2019) to our study. These graphs as they are built have too many false pipes compared to the real ones. This will cause a problem of unbalanced classes in the training step. The challenge here is to minimize the graph of possible connections without losing a large portion of the real pipes.

Our graph of possible connections is created following 7 steps:

1. Isolated pipes and duplicated nodes are removed from the Prades-Le-Lez and Ramonville-Saint-Agne networks.

2. The graph of connections is created as in (Chahinian et al., 2019), where the nodes are connected based on the Delaunay triangulation and a radius defined by the user (80 meters). The number of possible connections is almost 10,000 and 20,000 respectively for Prades-Le-Lez and Ramonville-Saint-Agne.

3. The graph of connections obtained from the previous step is directed. Since we do not consider the slope attribute, the directed graph is transformed into an undirected one. This reduces the number of possible connections by half, to almost 5000 and 10000 respectively for Prades-Le-Lez and Ramonville-Saint-Agne.

4. As indicated previously, public pipes are not buried under buildings. Hence, we removed all possible connections which fall within the buildings of each town. The buildings' databases are collected from IGN BD-TOPO©.

5. We have proceeded by reducing the number of possible connections for each node, by keeping only the n closest neighbors. In our case, when only the 6 shortest connections for each node are preserved, only 2% of the real pipes that were in the graph of possible connections are lost. The final graph of real connections thus contains 732 and 1,622 lines representing real pipes for Prades-Le-Lez and Ramonville-Saint-Agne. Yet, the new graph of possible connections is reduced to 1,828 and 4,449 lines respectively for Prades-Le-Lez and Ramonville-Saint-Agne.

6. The graph of possible connections is not used directly by the model. We transform the possible connections (potential pipes) into nodes and the edges are created when two possible connections share a node in the previous graph of connections.

7. The last step consists in defining the attributes for the nodes and the edges. We use only three attributes. The first one is the road attribute as defined in Section 3.4.1. The roads' databases are acquired from IGN BD-TOPO©. The second attribute is the length of the lines. The third attribute represent the edge attribute. They are the weights defined in the previous paragraph 3.4.2.1 which are based on the angles between the possibles connections.

We used WaGNN as shown in Figure 51. We learned from the graph of Ramonville-Saint-Agne and we tested our model on the graph of Prades-Le-Lez. We evaluated our results as follows:

- We compared WaGNN, using the precision (eq. 2.21) and the recall (eq. 2.20), to three different and popular approaches of GNNs, namely Graph-

SAGE (Hamilton et al., 2017) (with the LSTM aggregator since it produced better result than mean and pooling) as implemented in Deep Graph Library (DGL) (M. Wang et al., 2019), ChebNet (Defferrard et al., 2017) and TAGCN (Du et al., 2017) as implemented in Pytorch Geometric (Fey & Lenssen, 2019). After fine-tuning operations, we used 5 layers for each network with the outputs being respectively of size 64, 32, 16, 8 and 2. The best results with ChebNet were obtained by using $K = 4$ for the first 4 layers and $K = 1$ for the output layer.

- Although, there is a difference in the evaluation method, we compare our network to the work in (Chahinian et al., 2019) . We consider a line to be a real pipe only if the geometries match exactly (true positive). However, in (Chahinian et al., 2019), a line is considered to be a real pipe if it falls within a 5 meter buffer. Our method hence is more restrictive on positional errors than theirs.

3.4.3 Results and discussion

The metrics from Table 7, clearly indicate that WaGNN produces better results compared to the other GNN models. GrapheSAGE being the second best result, shows that learning an aggregation function is indeed more suitable for supervised learning, compared to ChebNet and TAGCN which focus on node embedding learning that is more suitable for semi-supervised situations as shown in our previous work (Belghaddar et al., 2021).

To compare WaGNN to the work in (Chahinian et al., 2019), we computed the four metrics defined in their study as follows :

- Completeness = $\frac{\text{TP}}{\text{TP+FN}}$
- Correctness = $\frac{\text{TP}}{\text{TP+FP}}$
- Quality = $\frac{\text{TP}}{\text{TP+FP+FN}}$
- Error = $\frac{\text{FN+FP}}{\text{RL}}$

Contrary to the equations 2.21 and 2.20, in which the number of pipes is used, these metrics are based on the length of the lines, where TP represents the length of true positives, FN the length of false negatives, FP the length of false positives and RL the total length of the 751 pipes of Prades-Le-Lez. Table 8 shows this comparison.

Although the study in (Chahinian et al., 2019) indicates better results compared to WaGNN when the slope is considered, we notice that, without the slope and despite a more restrictive method, we still achieved the same error(he only metric available).

Table 7 Prediction results on Prades-Le-Lez database using GNN models.

	GraphSAGE	ChebNet	TAGCN	WaGNN
Accuracy	0.84	0.82	0.80	**0.91**
Precision	0.84	0.82	0.80	**0.88**
Recall	0.75	0.72	0.68	**0.88**

Table 8 Comparison between WaGNN and (Chahinian et al., 2019) on Prades-Le-Lez database.

	(Chahinian et al., 2019)		WaGNN
	With slope	Without slope	Without slope
Completeness	0.92	–	0.86
Correctness	0.92	–	0.90
Quality	0.85	–	0.78
Error	0.16	0.22	0.22

To verify whether WaGNN performs better than the other GNN models in term of the length of the predicted pipes, we compared all the GNN models using the completeness, the correctness, the quality and the error. Table 9, shows that WaGNN outperformed all the GNN models.

Table 9 Length based comparison between the GNN models on Prades-Le-Lez database.

	GraphSAGE	ChebNet	TAGCN	WaGNN
Completeness	0.69	0.65	0.58	**0.86**
Correctness	0.86	0.83	0.80	**0.90**
Quality	0.62	0.57	0.50	**0.78**
Error	0.40	0.47	0.54	**0.22**

Figure 52, shows a map of the predicted pipes using our approach compared to the real map supplied by the network managers. Although the overall shape of the networks is similar, it is still necessary to carry out improvements. First, we notice that our model failed to produce a fully connected graph, since several pipes are isolated. Second, wastewater network maps do not include loops, since the main task

is to transport wastewater to a treatment plant. Hence, an ideal model should not produce such loops and our model does.

Despite these limitations our proposition can help managers to map their networks. Indeed, in situations were little information about the network is available, the produced map can be used as a basis to complementary survey operations in order to achieve a more accurate one. Also, given the required data: manholes, roads and building datasets, to our knowledge this approach can be considered as the best attempt to map a network without resorting to in-field or radar inspection.

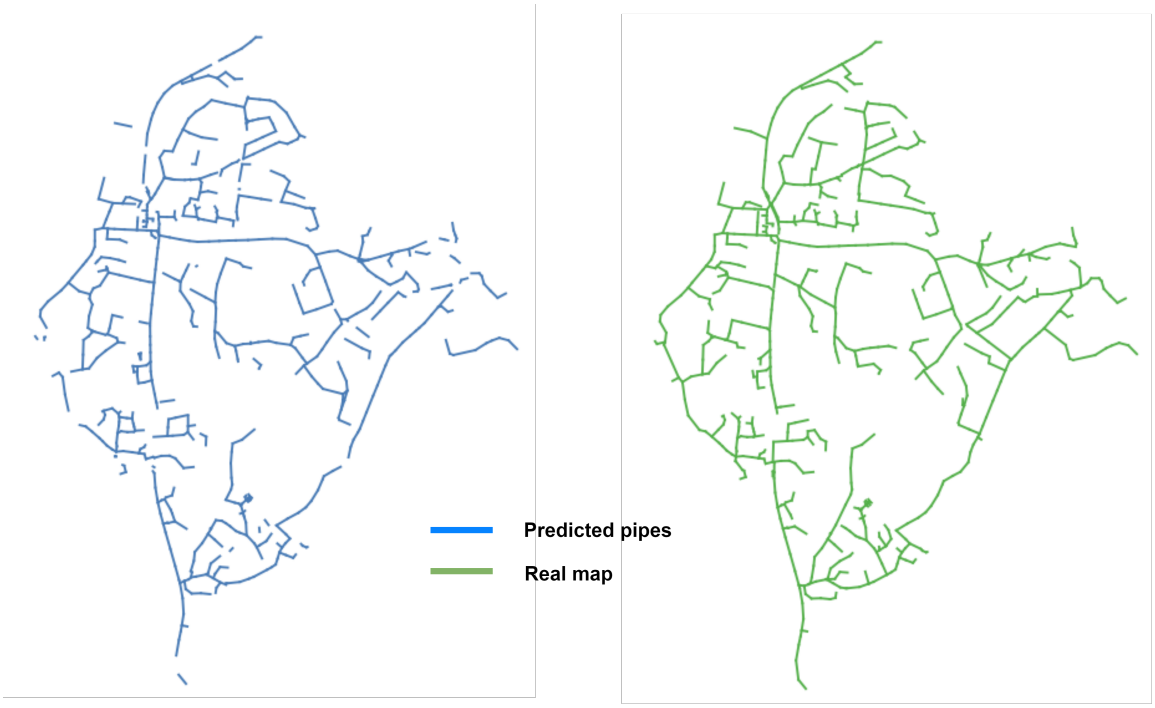


Figure 52 Predicted map versus real map.

3.4.4 Conclusions and perspectives

In this study we proposed a novel Graph Neural Network (WaGNN) specially dedicated to mapping wastewater networks. Given the positions of manhole covers, the task is to predict whether a pipe exists between two manholes. Inspired from the work in (Chahinian et al., 2019), the input to our network is a graph of possible connections between the manholes, which includes lines that represent the real pipes. We intentionally used only attributes that are often available in open access databases: roads and buildings datasets. Our GNN model is based on prior knowledge on the topology of wastewater networks which are organized as main branches composed of several pipes that ramify according to specific angle values. We trained the models on the database of Ramonville-Saint-Agne and the tests were carried out on the database of

Prades-Le-Lez. We compared our method to three different GNN approaches namely GrapheSAGE, ChebNet and TAGCN. Although there are major differences between our proposition and that of (Chahinian et al., 2019), we compared our results to their study. Also, despite having higher metric scores given the inputs and validation procedure in (Chahinian et al., 2019), we believe that our method is more practical, easy to execute and performs better. Having both recall and precision values of 0.88% suggests that we succeeded to learn to distinguish between the two classes even with unbalanced datasets. Thus, our proposition can be useful to managers in order to map their networks. Our method of course is not intended to replace field data and *in-situ* inspections when possible. However, it can represent a solid foundation to complementary survey operations, especially for municipalities with limited financial resources.

There is of course room for improvement and this work can be enhanced by adapting more domain knowledge. As shown in our results, WaGNN produces isolated pipes and loops which should not exist in a wastewater networks. Thus, to increase the accuracy of the produced map we aim at enhancing our network to address this issues. Multiple paths can be explored to achieve this goal, such as customized loss functions or a new aggregation operator. Contrary to cross entropy or other popular loss functions such as Mean Square Error, which are based only on the difference between the predictions and the actual labels, an ideal loss function for our application would be one that reflects the creation of the loops in the results. Of course, the difficulty here is whether we are able to find a formula that satisfies all the mathematical properties required by a loss function. Furthermore, at present, WaGNN doesn't have enough information to recognize that water should flow towards the outfall. Hence, we can imagine a new aggregation operator in which the message is transmitted differently depending on the position of the outfall. These issues will be addressed in our future work.

General conclusion

To provide continuous services for the population, the management of underground networks, particularly wastewater networks, involves expertise in different domains: economics, human resources, management, engineering, software development, etc. The literature review carried out in the framework of this PhD project showed that data related issues are part of the daily challenges the operators face. They may also cause incidents, traffic congestion, budget overruns, risk to environment and workers, etc. We focused this thesis on the challenges related to data, that we summarized in three characteristics:

- Data are collected from multiple sources.
- Data are heterogeneous.
- Data are imperfect: uncertain, imprecise and incomplete.

Given these characteristics, we have established that, in order to achieve more accurate and complete databases for wastewater networks, data fusion and imputation solutions are required. After studying these two different domains we set up four research questions:

1. In view of data fusion operations, how can wastewater data sources be modeled given their heterogeneous and imperfect nature?
2. How can object matching of wastewater networks be achieved and how can the imperfections be modelled?
3. How can the missing attribute values of wastewater networks be imputed and how can the structure of the graph be used for this purpose?
4. What domain knowledge or external information could be useful for both data fusion and imputation techniques and how can they be used?

From the start of this project, through the literature and the analysis of a dozen of databases that we managed to get our hands on, mainly from the local authorities of Montpellier Metropolis in France and the French open data, we deemed it necessary to study the data itself before addressing the problem of data fusion and imputation. Initially, we noticed a lack of abstraction related to both the components of the networks and their sources. To gain a deeper understanding of the networks' objects and their relationships, our first reflection was to propose a model for the components

of the networks and for the sources. We found that a very mature business model (COVADIS) was already established and recommended for the managers although it was not yet widely used by the managers. Since it was elaborated by multiple actors involved in the daily management of wastewater networks, we estimated that the model was sufficient to be adopted as an abstraction for the networks' components for our study. However the COVADIS model is designed for the french networks only and in other countries standards may vary. Nevertheless, although some information and the level of detail may differ between countries, the main information used to manage the networks (position, geometry, slope, elevation, diameter, type of material, installation/repairs date) remains similar. The second part of the abstraction was to encapsulate all the data sources in one entity to facilitate data fusion operations, which represent an answer to the first scientific question.

In chapter 1, we tackled this question by proposing a meta-model for the data sources of wastewater networks and its relationship with a given business model. Our contribution was motivated by two factors. First, the heterogeneity of the data sources and the diversity of the solutions to collect data: images, GPR data, GIS, etc, requires a unified framework in order to facilitate data fusion operations. Second, we chose a meta-model over a simple model since data collection methods could evolve over time and an exhaustive modelling of the sources is not a generic solution. Our proposition was inspired by the field of Big Data, since both domains' data sources share two important characteristics: the multitude of data sources and the heterogeneity of the data. Based on the difficulty of the pre-processing to extract relevant data from the sources, the latter were divided into three categories: structured, semi-structured and unstructured. We implemented our meta-model on the Moose platform and we showed using real examples that our proposition is generic and can be used to monitor and visualise data. It constitutes baseline to conduct fusion operations.

In chapter 2, we tackled the second and the fourth questions. After examining the related literature, we proposed a new process for matching the wastewater networks' spatial objects when collected from different sources. We aimed to model and take into consideration the imperfections of the data at the input and the output of the matching. This is intended to give the managers a more relevant and comprehensive view on the matching results. Identifying and understanding the origin of the common spatial imperfections in wastewater networks was imperative for all our decisions regarding the proposed steps for the object matching process. For instance, we addressed the problem of missing spatial objects, particularly pipes, by using the strokes as matching units in the first step, than conducting a partial matching later on. The similarity measures, which are the corner stone of any object matching ap-

proach, inherit the imperfections: uncertainty, imprecision and incompleteness of the data sources since they are defined from the available data (geographical, topological or attribute). These imperfections were essential to determine which data fusion technique to adopt among the available ones: Probabilities, Fuzzy sets, DS theory, etc. In fact, it was a great opportunity to study, understand and be aware of the syntactic and the semantic power of each theory as part of data fusion, a common problem shared by almost all domains that deal with data. DS theory turn out to be the more suitable for combining imperfect similarity measures related to wastewater networks. Indeed, it offered us the means to model the uncertainty of the attributes via the mass function, the imprecision/ignorance when the system doubted between two or more choices, the reliability of the sources using discounting operations and supporting the absence of some attribute values on the local combination step. We proposed an enhanced DS theory process for the combination step to reduce the conflict between the similarity measures. The results on synthetic data and real world data show that our process succeeded in matching the wastewater network's spatial object while keeping the conflict reasonable.

In chapter 3, we addressed the third and the fourth questions. We proposed three different contributions to deal with both spatial and non spatial missing data:

1. After identifying the minimum required attributes to run a hydraulic simulation. We investigated whether an automated process based on domain knowledge can first, enable us to run an end-to-end hydraulic simulation, second, provide comparable results to the ones established by hydraulic experts. The results of the experiments indicated that both goals were achieved although the set of algorithms that was developed for this task can be improved. Nevertheless, this study showed that even with limited available data, we can still estimate the networks' behaviour in different conditions.

2. To estimate the missing attribute values of wastewater networks, we compared popular traditional machine learning techniques (SVM, MLP and DT) to Graph Neural Network techniques (ChebNet, GCN, GraphSAGE and TAGCN). The aim of this study was to examine whether the structure of the network can help improve imputing the missing values. In this context, machine learning techniques were applied first to learn patterns from the available data and then to predict the missing values. Given that the available data vary between the networks, we defined two configurations of learning. In the first one we assumed severe conditions, where only the structure of the graph and portion of the attributes are available to im-

pute the missing values. As machine learning techniques require at least one attribute to learn, we constructed a new attribute based on domain knowledge. In the second configuration, we assumed that other correlated attributes to the targeted one can be available thus we added them to the ones in the first configuration. The results on real datasets showed that using the structure of the networks by aggregating information from the neighbors improves the imputation accuracy especially for high missing rates.

3. Contrary to the two previous propositions where the focus was on the attributes, the position of the buried utilities was the target here. We introduced a new Graph Neural Network (WaGNN) dedicated to predict whether a pipe exists between two manholes using very limited and often available attributes, namely road, length and building information. Inspired from the inductive method GraphSAGE, in WaGNN the message is received by each node depending on our knowledge about the usual topology of wastewater networks, namely the angles between the pipes. We showed using real datasets that for this task our method performs better than popular GNN models.

The results achieved in this thesis open the path towards new interesting research perspectives. In fact, our meta-model was tested in two occasions. The first one was when we instantiated a CSV file as a semi-structured data source and two unstructured data sources namely GSV and HR. The second was to achieve object matching on two structured data sources. Unfortunately, given the available data we couldn't investigate and evaluate the meta-model on more unstructured data sources, particularly signal based ones which are widely used data. Thus, the instantiation and the implementation of the necessary components are left for the future users and researchers to define. As for the relationship between the data sources and the business model, we used here the french model (COVADIS). Although, the meta-model was designed so the COVADIS can be replaced by other business models, a real experiment using other countries data would reveal whether adjustments are required. Moreover, the object matching process was an opportunity to verify the expressiveness of the meta-model in terms of imperfections. Indeed, the necessary information at source level and attribute level were sufficient for the process that we proposed. However, one may decide to apply other approaches such probability theory, thus the researchers must define any supplementary data required for such process. For the spatial matching itself, as indicated in the COVADIS, pipes and nodes can be divided into multiple types. In the current state of our proposition, this aspect is not considered. Hence, future research may be directed towards a more detailed object

matching procedure. DS-theory offer tools for such problems, the refinement or the generic extension operator (Appriou, 2014) is one. Nevertheless, other mathematical tools such as Bayesian inference can also be investigated.

The third chapter uses domain knowledge and external information the most. We believe that more information could be explored in future studies to enhance the imputation results. For instance, the type of buildings near a pipe may give an indication about the diameter of a pipe and the input discharge, or public reports can reveal the type of material imposed by the decision makers during the construction of the network. A good example of data collection about wastewater regulations and interventions is the one in (Chahinian et al., 2021), where information are extracted automatically from the web. Nevertheless, the most important addition regarding data imputation was showing the effect of considering the topology of the networks on the imputation results. We demonstrated in two different tasks the power of the GNN models to exploit the relationships between the objects to learn hidden patterns. However, in our case, the imputation of attributes using the popular GNNs was limited to discrete values (diameter and type of material). The slope attribute – which is the only necessary attribute for hydraulic simulation – having continuous real values, was not considered. Hence, future studies can be directed to use or develop a GNN model using a semi-supervised regression for slope estimation. We are certain that more advanced GNN models, driven by the specification of the addressed problem, can be developed for different applications related to wastewater networks especially and data represented by a graph, such as roads, in general. Indeed, we showed that although popular machine learning models may perform well in benchmark datasets, they can be drastically enhanced when guided by a profound understanding of the nature of the task at hand, translated in our case by sending different messages between the nodes according to the angle attribute. However, the mapped network was characterized by clear issues that can be avoided: the loops between the pipes and isolated pipes from the outfall. In future works, researches can address these issues, for instance by using a dedicated loss function to penalize such relationships in the final map.

The distinct solutions that we proposed in this research topic demonstrate our intention to cover as many aspects related to the data management of a wastewater network as possible. As stated previously, our goal was not to replace in field inspection, but to offer an alternative and complementary tools for the managers to enhance the quality of their databases when exhaustive inspection of the buried utilities is not feasible. Our propositions covered all the aspects involved in the mapping of the networks: data management, data collection, data fusion and data imputation. When a network is already mapped correctly, the managers may dedicate more time

and efforts to prioritize urgent maintenance operations and anticipate leaks. As long term perspective, we trust that the topology of the networks would be of great help to address this problem. Beyond wastewater networks, we hope that the extensive use of the domain knowledge and the structure of the graph in the different tasks, could inspire and guide researchers in their future studies regarding underground networks in general.

Appendix

Résumé en Français:

Fusion de données pour la cartographie de réseaux urbains: application aux réseaux d'assainissement

Table des matières

Introduction

Aujourd'hui, plus de la moitié de la population vit dans les villes et dans 35 ans ce sera plus de 70% [1]. Pour répondre aux différents besoins de cette concentration urbaine, partout dans le monde des réseaux souterrains permettent chaque jour et d'une manière transparente de transporter, distribuer et fournir l'ensemble des éléments nécessaires à notre vie quotidienne (Gaz, eau, électricité, internet, etc.). Or, ces réseaux, depuis leur installation, subissent fréquemment des opérations de réparation et d'expansion, ce qui rend leur gestion complexe, coûteuse et fastidieuse.

Pour éviter de creuser aux mauvais endroits ou d'endommager des conduites en places, les excavateurs doivent demander et collecter toutes les informations identifiant la position et la nature des objets enterrés [2]. Cette opération de repérage, de localisation et d'extraction d'informations précises sur les réseaux est contrainte à plusieurs obstacles, en voici quelques uns :

- Les sociétés et les entreprises qui participent à la maintenance des réseaux souterrains sont multiples et changent régulièrement. Par conséquent, les plans disponibles sont parfois incohérents, incomplets et contradictoires.
- Les documents concernant les interventions, lorsqu'ils existent, datent d'époques différentes, ce qui rend leur exploitation difficile.
- Les données concernant les objets constituants les réseaux souterrains (regards, conduite, etc.) sont de formats et de types différents : images, GIS, fichiers texte, tableaux, BD, etc.

Cette liste n'est pas exhaustive et probablement les spécialistes et les intervenants pourront nous révéler d'autres défis. Ce qui est certain, c'est que l'inexactitude, l'incomplétude et l'incertitude des sources de données se reflètent de manière critique sur la gestion des réseaux souterrains, et divers sont les exemples qui illustrent cette réalité. Les travaux de réparation et de mise à jour impliquent généralement le blocage des routes pendant des heures voire des jours. Chaque année en France plus de 100 000 endommagements de réseaux se produisent lors des travaux à leur proximité [3]. Du point de vue économique, les coûts associés aux travaux publics de la réparation et la mise à jour des réseaux enterrés au-dessous des routes est en croissance rapide (7B£/ans en GB) [4].

La collecte d'informations imparfaites, hétérogènes et multi-sources soulève inévitablement le problème de fusion de toutes ces données. Pour résoudre ce problème nous nous sommes posé deux questions principales :

- Comment fusionner ou combiner des données obtenues à partir de plusieurs sources dans le contexte des réseaux d'assainissement ?
- Quelles solutions proposer pour les données manquantes des réseaux d'assainissement ?

À partir de ces deux questions et à travers l'étude des domaines de fusion de données, d'estimation des données manquantes, nous avons identifié 4 questions scientifiques pour améliorer les cartes des réseaux d'assainissement :

- Dans la perspective de conduite des opérations de fusion de données, comment les sources de données sur les réseaux d'assainissement peuvent-elles être modélisées compte tenu de leur nature hétérogène et imparfaite ?

- Comment réaliser l'appariement d'objets des réseaux d'assainissement et comment modéliser leurs imperfections ?
- Comment estimer les données manquantes des réseaux d'assainissement et comment la structure des réseaux peut-elle être exploitée à cette fin ?
- Quelles connaissances métiers pourraient être utilisées à la fois pour la fusion des données et l'estimation des données manquantes ?

Les réponses à ces questions sont présentées sous forme de trois contributions dans les 3 chapitres de ce document :

- Le chapitre 1 décrit le méta-modèle proposé pour les sources de données des réseaux d'assainissement, prévu pour faciliter des opérations de fusion de données.
- Le chapitre 2 présente notre proposition d'appariement des objets, précisément les conduites des réseaux d'assainissement. Cette contribution se base sur la théorie de Dempster-Shafer pour modéliser les imperfections des sources de données.
- Le chapitre 3 se focalise sur la complétion des données manquantes en se basant sur la topologie des réseaux d'assainissement.

Chapitre 1

Meta-modélisation des données des réseaux d'assainissement

1.1 Introduction

Le développent rapide des systèmes d'informations durant ces trois dernières décennies a rendu l'accès facile à des nouveaux moyens de collecte et de stockage de données. Avant toute opération sur les réseaux d'assainissement, les intervenants collectent toutes les informations nécessaires pour éviter de creuser aux mauvais endroits et endommager des installations en place. Parmi les moyens utilisés pour acquérir ces informations : les données radar, les cartes numériques et analogiques disponibles, les rapports d'interventions, etc. Cependant, cette diversité des sources et l'hétérogénéité de leurs structures et formats rendent aujourd'hui leur exploitation difficile du fait de leurs écarts conceptuel, technique et sémantique.

Pour prendre des décisions éclairées, les données collectées sont analysées, comparées et combinées. Cependant, en plus de l'hétérogénéité des sources, les imperfections des données (incertitude, imprécision et incomplétude) font surgir des problèmes d'incohérences et d'incertitude. Ainsi, cette tâche de combinaison devient complexe et chronophage.

La combinaison de données multi-sources nécessite également un modèle de données unifié pour permettre la centralisation, la mise à jour, l'archivage et le suivi de ces données. En effet, nous avons analysé des bases de données numériques liées aux réseaux d'assainissement pour comprendre la sémantique de leurs données, leurs relations et déterminer leurs différences. Les modèles de données associés, lorsqu'ils existent, étant rarement accessibles au public, nous avons procédé en les inférant à partir des données. Dans notre étude, nous avons utilisé les données fournies par des sources fiables, notamment les données publiques en France [5]. Parmi les fournisseurs figurent la métropole de Montpellier, la communauté urbaine du Sud-Est de Toulouse Sicoval, Data Angers, et la région des Pays de la Loire.

Suite à l'étude de ces différentes sources, nous avons identifié les contraintes suivantes :

- Les modèles de données adoptés par les opérateurs sont différents. Par conséquent, l'échange et la réutilisation des données sont difficiles.

- Les modèles ne sont pas conformes aux règles et normes de conception et de modélisation informatiques.
- Les attributs fournis par les propriétaires de données sont liés à leurs domaines d'activité. Par exemple, une entreprise spécialisée en modélisation hydraulique fournit des informations précises sur le débit de l'eau dans une canalisation, alors que ces mêmes informations sont généralement absentes des données provenant d'une autre entité experte en réparation d'ouvrages.
- L'historique des interventions, nécessaire à l'anticipation des réparations, est rarement pris en compte dans ces modèles.

Un modèle générique pour les données métiers et les sources de données est alors nécessaire pour résoudre les problèmes dus à l'écart conceptuel et sémantique entre les modèles de cartes numériques existants et pour permettre la mise en œuvre d'une approche efficace de fusion de données. Dans le but de proposer un tel modèle nous avons commencé d'abord par un état de l'art sur les travaux en relation avec les réseaux d'assainissement, puis une étude des concepts de modélisation en général.

1.2 Etat de l'art

1.2.1 Travaux connexes aux modèles de données métiers des réseaux d'assainissement

Les divers travaux visant à aider les opérateurs à collecter et améliorer les données sur les réseaux d'assainissement, tels que ceux qui se basent sur les images [6] ou les radars [2], ne sont pas fournis avec un modèle de données pour répertorier les données cibles ainsi que leurs relations. De plus, dans la littérature peu d'études se sont intéressées à la modélisation des données des réseaux d'assainissement. Nous n'avons identifié que deux travaux : dans [7], afin de construire la carte numérique du réseau d'assainissement d'une commune en Algérie et de contribuer à sa gestion efficace, un modèle de données métier a été proposé à partir de l'inventaire des données disponibles. À l'initiative de la Région Aquitaine et du Groupement d'intérêt public "Aménagement du territoire et gestion des risques", la Commission de validation des données pour l'information spatialisée (COVADIS) [8] a publié un standard de données pour les réseaux d'eau potable et d'assainissement collectif.

Le modèle COVADIS a été établi pour faciliter la communication entre les intervenants de la Région Aquitaine. Ensuite, il a été généralisé au niveau national en France et aujourd'hui les communes, les municipalités et les responsables des réseaux sont appelés à respecter ce standard. Pour ces raisons, nous avons adopté ce modèle de données métiers pour les réseaux d'assainissement pour notre étude. Nous présenterons dans ce qui suit la partie du diagramme de classe COVADIS relative à l'assainissement (Figure 1.1).

Le standard de données COVADIS est composé de 4 classes principales :

- Nœuds : représentés spatialement par des points, ils illustrent des appareils (vanne, compteur, etc.), des regards ou les lieux de jonctions entre les conduites.
- Canalisations : représentées spatialement par des lignes, elles sont classées en plusieurs catégories : eaux usées, eaux pluviales, etc. Chaque canalisation possède deux Nœuds d'extrémités.

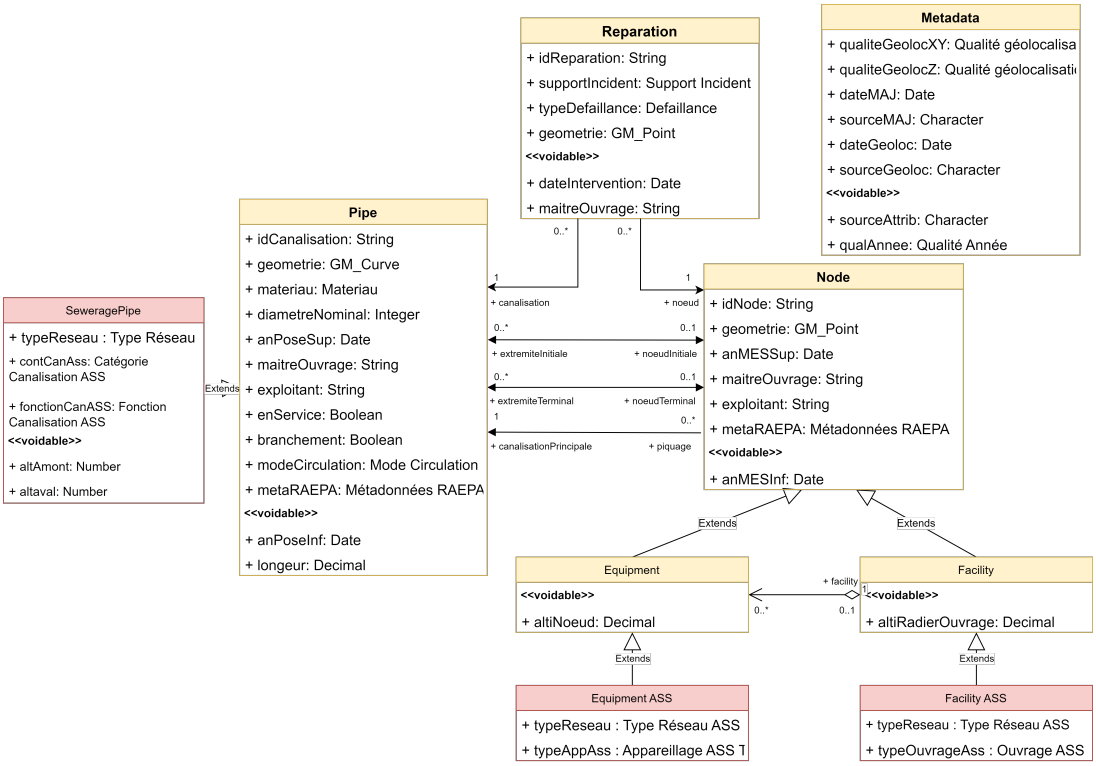


FIGURE 1.1 : COVADIS : le modèle métier des réseaux d'assainissement [8].

- Réparations : représentées spatialement par des points, elles font référence à des interventions effectuées dans des Nœuds ou des Canalisations.
- Métadonnées : sont des données utilisées pour qualifier les informations des classes Nœud et Canalisation ; À savoir, le nom de la source, la date de la dernière mise à jour, la fiabilité de l'année d'installation et la qualité de la géolocalisation.

1.2.2 Réseaux d'assainissement et Big Data

Le domaine du Big data et celui des réseaux d'assainissement partagent deux caractéristiques importantes :

- La multitude des sources de données.
- L'hétérogénéité des sources de données.

Les travaux de recherche dans le domaine des réseaux souterrains sont majoritairement confidentiels [9]. Par conséquent, le nombre de publications liées au Big Data est plus important que celui sur les réseaux d'assainissement. À notre connaissance il n'existe aucun modèle de données pour les sources de données pour les réseaux souterrains. Pour combler ce manque et puisque les deux caractères de multitude et d'hétérogénéité des sources ont été déjà examinés en Big Data (par exemple dans [10] ou [11]), nous avons choisi de nous inspirer des solutions proposées dans ce domaine.

Les systèmes et plates-formes Big Data disponibles ne sont pas similaires, car ils proviennent de multiples fournisseurs dont la vision n'est pas identique. Dans [12], en utilisant l'ingénierie dirigée par les modèles, les auteurs proposent un méta-modèle indépendant des plate-formes pour décrire les structures des sources de données impliquées dans l'alimentation de ces grands volumes de données, permettant ainsi aux programmeurs de créer des applications compatibles avec divers produits. Étant donné que les Big Data comprennent trois formats hétérogènes : Non structuré, Semi-structuré et Structuré [13], les auteurs proposent une méta-modélisation des trois types de sources de données comme suit :

- Structurées : données dont l'ensemble des valeurs possibles est déterminé et connu à l'avance, comme les bases de données relationnelles.
- Semi-structurées : données qui n'ont pas été organisées dans une entité spécialisée. Cependant, elles contiennent des méta-données, qui aident à leur exploitation, par exemple les e-mails.
- Non structurées : données représentées ou stockées sans format prédéfini.

1.2.3 Méta-Modèles

Plusieurs définitions ont été proposées pour « méta-modèle ». Selon [14], un méta-modèle est un modèle utilisé pour modéliser la modélisation elle-même. Un méta-modèle est un modèle qui définit la structure d'un langage de modélisation [15]. L'idée de base d'un méta-modèle est d'identifier les concepts généraux dans un domaine ou pour une problématique donnée et les relations utilisées pour décrire les modèles [16]. Cette généralité que nous cherchons à satisfaire dans notre proposition est l'un des

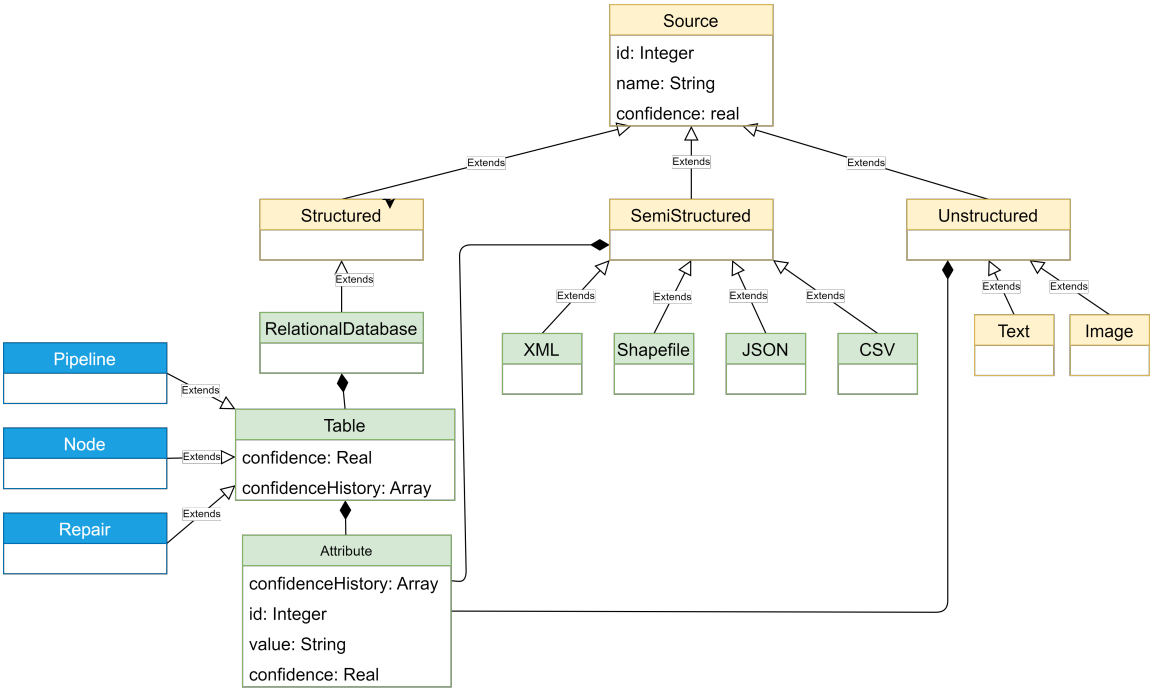


FIGURE 1.2 : Méta-modèle pour les sources de données des réseaux d'assainissement

éléments majeurs qui ont rendu la méta-modélisation l'une des approches les plus importantes pour la modélisation. Les instances d'un méta-modèle sont des modèles qui doivent satisfaire les spécifications du méta-modèle. Ainsi, elles permettent de modéliser les systèmes cibles de manière cohérente et homogène. Les méta-modèles sont utilisés dans plusieurs domaines, par exemple dans [17], les propriétés associées à un logiciel en exécution sont modélisées par un méta-modèle. Les auteurs dans [18], proposent un méta-modèle pour la surveillance des systèmes cyber-physiques (CPS), en particulier les capteurs et les réseaux d'actionneurs qui nécessitent des données valides et une bonne coordination entre les capteurs et des actionneurs pour leur fonctionnement.

1.3 Contribution

Les sources de données évoluent dans le temps, par conséquent une modélisation exhaustive des sources n'est pas générique. Nous proposons ici un méta-modèle pour les sources de données des réseaux d'assainissement. Ce méta-modèle devrait servir un point de départ pour effectuer des opérations de fusion et d'intégration des données, tout en prenant en considération les imperfections des sources. La Figure 1.2 illustre notre méta-modèle. Pour une meilleure compréhension, nous présenterons deux points de vues de ce méta-modèle.

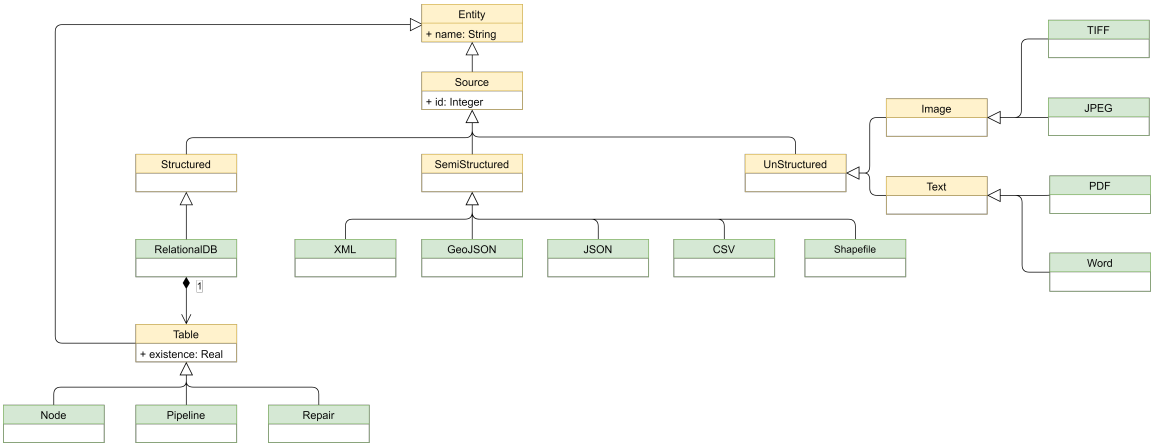


FIGURE 1.3 : Point de vue des sources de données

1.3.1 Point de vue des sources de données.

Nous résumons dans la Figure 1.3, le point de vue des sources de données où l'élément principal « source » caractérise toute entité capable de fournir des données, des informations ou des connaissances sur les réseaux d'assainissement. La distinction entre les sources de données est définie comme suit :

- Sources non-structurées : dont les formats nécessitent un pré-traitement important avant d'extraire des données métiers sur un réseau. Le pré-traitement de ces sources produit généralement des données semi-structurées sur les réseaux. Par exemple, fichier CSV pour l'emplacement des regards d'égouts détectés à partir des images.

- Sources semi-structurées : dont les formats nécessitent un pré-traitement simple avant d'extraire des données métiers sur un réseau. Par exemple, analyser des fichiers CSV ou XML.

- Sources structurées : représentent des bases de données relationnelles qui fournissent directement des données métier sur un réseau. Généralement, ces données sont fournies par les gestionnaires officiels des réseaux d'assainissement.

1.3.2 Point de vue de la confiance associée aux données

L'imperfection des données est considérée dans ce point de vue en associant des attributs de confiance à chaque entité (Figure 1.4). Cela implique :

- Chaque source a une valeur de confiance qui indique la fiabilité ou la certitude des informations qu'elle fournit. Cette métrique peut être modélisée par les différents outils mathématiques disponibles, comme les probabilités.

- Pour les composants ou les objets des réseaux d'assainissement, cette valeur représente l'incertitude sur leur existence. En effet, il est possible, par exemple, qu'une canalisation ou un regard soit représenté sur une carte par erreur.

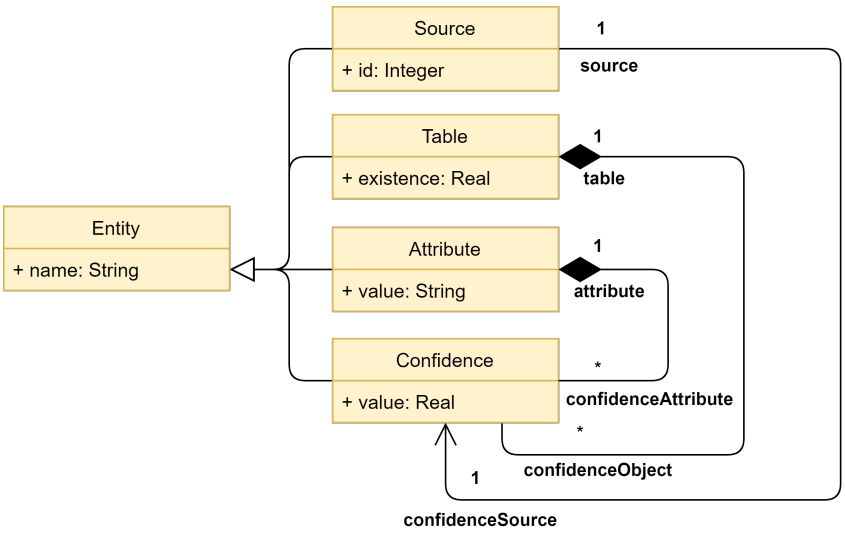


FIGURE 1.4 : Point de vue de la confiance associée aux données.

- Pour les attributs, la confiance est liée à celle de la source de données qui les fournit.

1.4 Cas d'utilisation

À travers notre méta-modèle, nous souhaitons réaliser la fusion des données provenant de différentes sources. Pour atteindre cet objectif, quelle que soit l'architecture adoptée, centralisée ou distribuée, les données hétérogènes collectées à partir de diverses sources doivent être agrégées. La Figure 1.5, décrit les étapes pour aboutir à cette agrégation. La première étape concerne les sources non-structurées où l'extraction des données nécessite un pré-traitement important. Généralement, cette étape produit des données organisées dans un format semi-structuré. La deuxième étape consiste à appliquer les transformations nécessaires pour adapter les sources structurées et semi-structurées au modèle métier cible, tel que le standard COVADIS.

Pour illustrer ce processus, nous avons mené une expérience d'agrégation de données sur le réseau d'assainissement de Prades-le-Lez, commune située en périphérie de la ville de Montpellier en France, provenant de 3 sources différentes. La première source est la base de données de Montpellier Méditerrannée Métropole, stockée dans un format CSV, qui est une source semi-structurée que l'on note la source 3M. Les deux autres sources sont une image haute résolution et des images Google Street View, qui sont toutes les deux des sources non structurées que nous notons respectivement sources HR et GSV. Les objets détectés à partir de la source HR ont été enregistrés dans un fichier XML, et ceux de la source GSV dans un format CSV.

Nous avons instancié notre méta-modèle en utilisant la plate-forme Moose. Moose [19, 20] est une plate-forme open source pour l'analyse des logiciels et des données, actuellement basée sur Pharo [21] un langage de programmation purement orienté objet. La plate-forme offre plusieurs fonctionnalités :

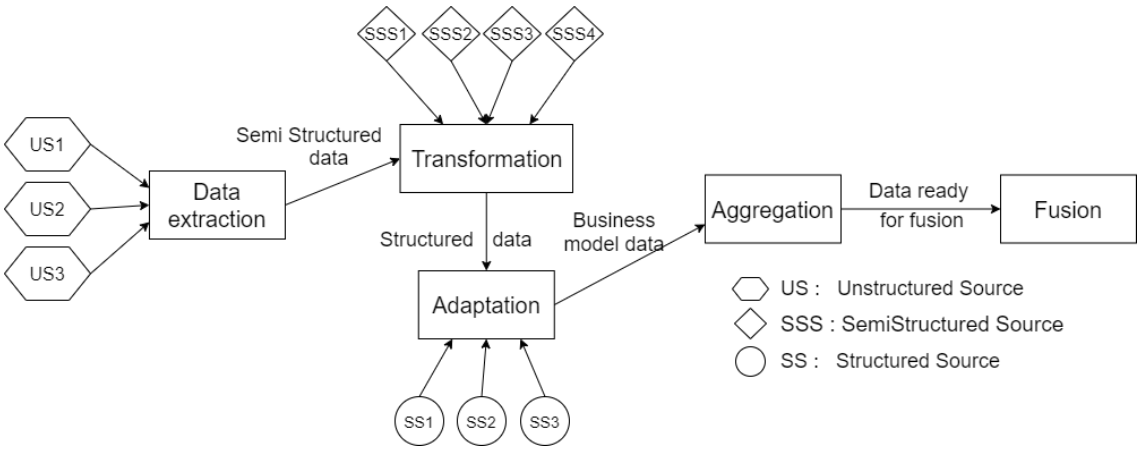


FIGURE 1.5 : Les étapes principales envers une fusion des données.

- Importation et méta-méta-modélisation, puisque la première étape du processus d'analyse est la génération d'un modèle pour un système cible donné.
- Parsing, qui fournit une interface fluide pour une construction facile.
- Une interface riche pour interroger les modèles.
- Visualisation à travers des graphiques et des tableaux.
- Navigation, qui permet à l'analyste de parcourir n'importe quel modèle.

TABLE 1.1 : Exemples de requêtes sur les trois sources.

(a) Les valeurs de confiances associées aux sources.

Sources	CSV_3M_source	XML_GSV_source	CSV_HR_source
Confidence	0.90	0.60	0.75

(b) Les valeurs de confiances associées à l'attribut position.

CSV_3M_source			XML_GSV_source			CSV_HR_source		
ID	Position	Confidence	ID	Position	Confidence	ID	Position	Confidence
613	3.8632613,43.6916825	0.90	78	3.8632613,43.6916825	0.75	12	3.8632613,43.6916825	0.76
622	3.8642639,43.7010732	0.90	80	3.8642639,43.7010732	0.58	52	3.8642639,43.7010732	0.80
623	3.8660806,43.6890539	0.90	45	3.8660806,43.6890539	0.98	32	3.8660806,43.6890539	0.84
602	3.8653489,43.6939405	0.90	14	3.8653489,43.6939405	0.96			
617	3.8629708,43.6884132	0.90	19	3.8629708,43.6884132	0.94			
						66	3.8734563,43.6932746	0.82
						207	3.8646406,43.7005578	0.80

Pour agréger des données provenant de plusieurs sources dans Moose, un coordinateur est nécessaire. Nous avons défini ModularIDGSMToMsgMonitoringConverter, qui contient les différents convertisseurs prescrits pour chaque source de données en tant que module. Chaque convertisseur contient un Importeur, un Parseur, un Manager du modèle et un Visualiseur. Le rôle de ce coordinateur est de choisir automatiquement pour chaque source de données, le convertisseur correspondant. Le coordinateur reçoit séquentiellement des données provenant de différentes sources de données, puis pour chaque

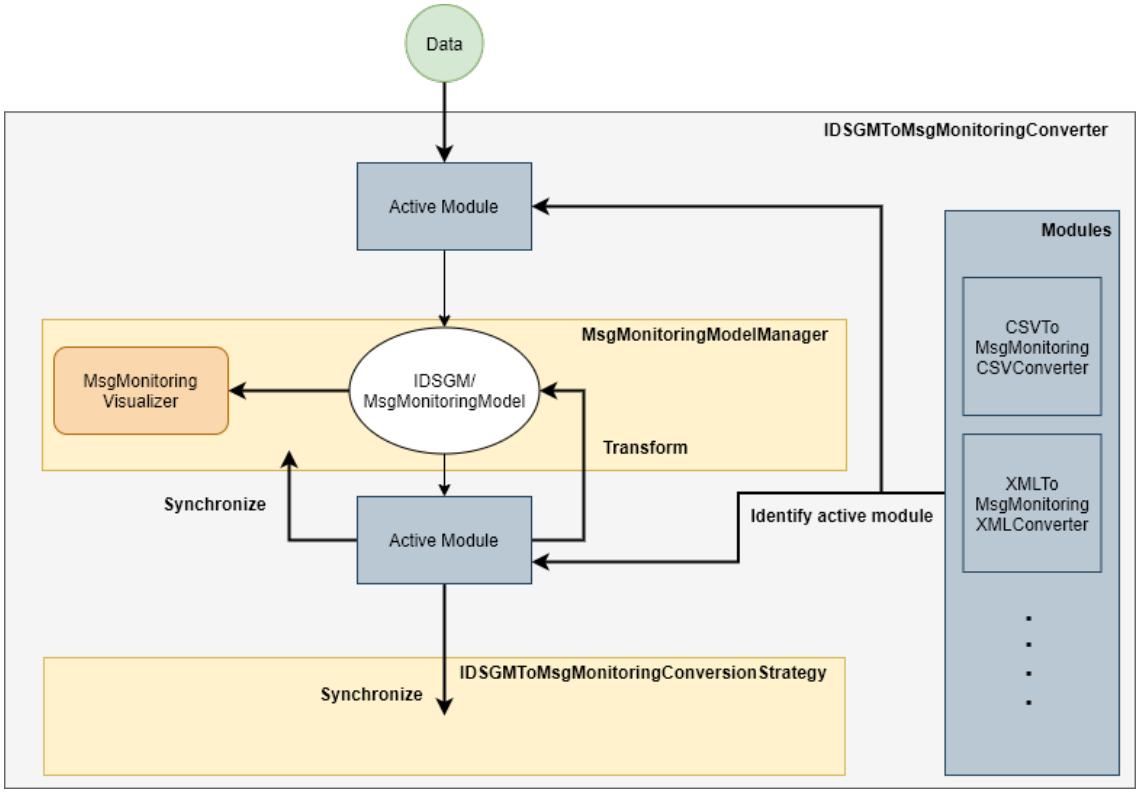


FIGURE 1.6 : Le processus d'agrégation de plusieurs sources de données(CSV et XML)

itération, le module associé à la source de données est identifié et utilisé. La Figure 1.6 illustre ce processus. Le Tableau 1.1 montre des exemples de requêtes que nous avons effectuées pour interroger les trois sources. Le Tableau 1.1a indique la fiabilité de chaque source et le Tableau 1.1b indique les positions des objets détectés ainsi que les valeurs de confiances associées.

1.5 Conclusion

Dans ce chapitre, nous avons proposé un méta-modèle pour les sources de données des réseaux d'assainissement inspiré du domaine du Big Data. L'objectif principal était de fusionner les données des réseaux d'assainissements, afin de mettre à la disposition des décideurs un ensemble de données plus précis et complet. Notre proposition a pris en compte trois aspects importants : i) la structure des sources de données (structurées, semi-structurées et non structurées), ii) les confiances associées à plusieurs niveaux et iii) la généricité liée au lien avec les modèles métier.

Comme première étape vers la fusion des données des réseaux d'assainissement, nous avons implémenté cette étape dans Moose, une plateforme de logiciels et d'analyse de données. Nous avons employé des exemples concrets pour montrer la genérécité de notre méta-modèle. Nous avons utilisé une source semi-structurée et deux sources non structurées avec les processus d'extraction de données appropriés. Les résultats ont permis le recensement, la visualisation et l'analyse des données agrégées. Cette contribution a été publiée dans [22].

Chapitre 2

Appariement d'objets en se basant sur la théorie DS

2.1 Introduction

Dans la littérature, deux concepts liés à la combinaison des données spatiales émanant de sources différentes sont identifiés : la conflation et l'intégration des cartes. L'intégration des données est définie dans [23] comme le processus d'unification des sources de données dans un cadre unique. Elle prend en entrée des bases de données (les schémas de données et leurs instances) et elle génère un schéma unifié et un mapping pour pouvoir extraire les instances à partir des anciennes bases de données. La conflation est définie par [24] comme le processus de création d'un nouvel ensemble de données, basé sur deux ou plusieurs jeux de données différents couvrant la même zone. Lors de l'intégration/fusion des données spatiales, l'étape la plus importante et complexe est celle de l'appariement d'objets [25, 26, 27], elle consiste à déterminer les objets correspondants dans différents ensembles de données, qui représentent des entités analogues dans le monde réel [24].

L'appariement d'objets est difficile non seulement en raison des différences en termes de résolution, de schémas, de temporalités et de représentations entre les sources, mais aussi et surtout à cause des imperfections des données telles que l'incomplétude, la distorsion et l'imprécision. Gérer les imperfections dans l'appariement d'objets revient à répondre à deux questions : i) Comment traiter ou modéliser les imperfections ? ii) Comment refléter ces imperfections sur les résultats d'appariements ? Pour traiter les imperfections, les objets de différents jeux de données sont comparés selon plusieurs critères tels que la position ou la forme et la correspondance est décidée en les combinant. En fonction des imperfections des données étudiées, la contribution de chaque critère change. Par exemple, lorsqu'un jeux de données souffre d'importantes distorsions, la contribution d'un critère basé sur la topologie est plus importante que celle d'un autre basé sur la distance. Traditionnellement, la contribution de chaque critère est modélisée par un poids, et l'incertitude de l'appariement est calculée comme la moyenne pondérée des critères choisis. Outre cette approche de base, certaines études comme [24], ont utilisé la théorie des probabilités pour produire une incertitude plus représentative des résultats d'appariements. La théorie de Dempster-Shafer (DS), qui permet de modéliser l'incertitude et l'imprécision [28], a été

utilisée dans [29, 30, 31] pour obtenir la correspondance d'objets sur des données de routes et de points d'intérêts.

Dans ce chapitre nous proposons un processus pour l'appariement des objets des réseaux d'assainissement. Notre proposition prends en compte les imperfections des sources en s'appuyant sur la théorie de Dempster-Shafer. La suite de ce chapitre est organisée comme suit : la section 2.2 présente l'appariement des objets et ses étapes principales. La section 2.3 est une brève introduction des concepts de base de la théorie de DS. Notre processus d'appariement est décrit dans la section 2.4. Les résultats sont présentés dans la section 2.5 et les conclusions de ce travail dans la section 2.6.

2.2 Appariement d'objets

Dans la littérature plusieurs définitions ont été proposées. Pour [27], l'appariement des objets est l'identification d'objets correspondants dans diverses sources de données. Pour effectuer l'appariement d'objets, nous avons besoin de deux outils :

- Une ou plusieurs mesures de similarités entre les objets.
- Une méthode qui utilise ces mesures, selon un processus bien défini, pour prendre des décisions.

2.2.1 Les mesures de similarités

Les mesures de similarités sont utilisées pour comparer les instances [32]. Elles représentent les critères sur lesquels se base un appariement [33]. La mesure de similarité la plus intuitive est celle qui s'appuie sur la position : nous supposons que deux objets sont homologues lorsqu'ils sont proches en termes de distance. La distance la plus utilisée est l'euclidienne [27, 34, 35]. Or, lorsque nous avons des sources de données qui ont un écart ou des distorsions importantes et non uniformes, les objets qui se correspondent ne sont pas forcément les plus proches. Par conséquent, il serait judicieux d'utiliser des attributs, le nombre de voisins, la topologie, etc.

Dans la littérature, plusieurs mesures de similarités/distances ont été proposées. Les concepts de distance et de mesure sont complémentaires car généralement, l'un est défini par rapport à l'autre [36, 35]. Par exemple, dans [35], transformer une distance en mesure consiste d'abord à normaliser la valeur de distance sur l'intervalle [0,1] puis à garder le complément à 1.

Historiquement, selon [26, 37, 38], la majorité des premières théories et méthodes sur l'appariement des objets, plus spécifiquement dans le cadre de la conflation des objets, proviennent des travaux du bureau de Census aux Etats-Unis entre 1983 et 1985 [32, 39]. Depuis, un grand nombre de mesures ont été proposées. Nous les classifions en 3 catégories :

- Géométriques : position, angle, longueur, etc.
- Topologiques : degré d'un nœud, directions des lignes, contexte géographique, etc.
- Attributaires : nom, adresse, etc.

2.2.2 Les méthodes d'appariements

Dans la littérature un nombre important d'approches pour apparier les objets a été proposé. Ces approches peuvent être classifiées selon plusieurs critères. Dans cette section nous allons les présenter et les classifier brièvement selon deux axes.

1- Classification selon les contraintes supportées

Le choix des méthodes appropriées parmi les propositions disponibles nécessite la satisfaction des contraintes imposées par les jeux de données cibles. Chaque approche proposée dans la littérature peut être caractérisée par les quatre contraintes suivantes :

- Type de représentation : points, lignes ou polygones.
- Les échelles à intégrer : similaires ou différentes.
- Cardinalités : $(1:1)$, $(1:0)$, $(0:1)$, $(1:m)$, $(N:m)$ ou $(N:1)$.
- Nombre de sources : 2 ou N.

2- Classification selon les étapes d'appariement

Comme indiqué précédemment, toutes les méthodes d'appariement utilisent des mesures de similarité pour identifier les candidats possibles à apparier. À l'exception des cas comme dans [34], où la distance entre les nœuds est la seule mesure disponible, la plupart des méthodes d'appariement utilisent diverses mesures. Par conséquent, les méthodes se distinguent par la façon dont elles définissent, utilisent et combinent ces mesures. D'abord, nous distinguons les méthodes indépendantes des mesures (c'est-à-dire que les distances ne sont que des paramètres à modifier), et celles dans lesquelles les mesures font partie du processus d'appariement, comme dans [40], où les auteurs utilisent la classification de Horton et des nœuds pré-appariés pour faire correspondre des réseaux hydrographiques. Aussi, nous pouvons distinguer entre les méthodes monodirectionnelles où la correspondance est faite dans un sens à partir d'un jeux de données *de référence* vers un autre jeu *cible* [25]. Et les méthodes bidirectionnelles où les jeux de données sont utilisés simultanément comme référence et cible [37]. Le processus d'appariement peut être résumé en 3 étapes principales :

1. Sélection des candidats : consiste sélectionner les objets considérés proches. Par conséquent, réduire l'espace de recherche pour un objet (point, ligne ou polygone) pour lequel nous cherchons un(des) correspondant(s). Cette opération est effectuée en utilisant un « buffer » (zone tampon) autour des objets, des filtres/seuils appliqués aux mesures, ou les deux. Un objet peut être considéré seul ou pris dans un groupement. Par exemple, les rivières, les routes et les réseaux souterrains se caractérisent par des branches principales qui se ramifient en plusieurs sous-branches. Pour considérer cet aspect hiérarchique, dans [41, 40, 42], les auteurs proposent d'utiliser le concept de « stroke » comme unité de correspondance, où un stroke décrit une « bonne continuité » entre des lignes, i.e. qui ne dépassent pas un degré de déviation choisi. Par conséquent, lorsque des nœuds sont présents dans un ensemble de données et absents dans les autres, l'utilisation des strokes capture la structure globale des réseaux et aide à obtenir une correspondance indépendamment des nœuds manquants.

2. Combinaison des mesures de similarités : sur la base des mesures utilisées dans la sélection des candidats ou en définissant de nouveaux critères, un score de similarité est calculé entre les objets. Généralement, le score est la somme du produit des mesures par leur poids. Dans [31, 30, 29], les auteurs utilisent la théorie de Dempster-Shafer comme formalisme pour combiner les mesures de similarités. Dans ce cas, les mesures sont transformées en fonctions de masses, ensuite fusionnées, pour enfin prendre les décisions. Comparée à d'autres théories comme celle des probabilités qui permet de modéliser seulement l'incertitude, la théorie de DS permet de modéliser l'incertitude et l'imprécision des données. En plus, elle offre un ensemble d'outils mathématiques pour combiner et gérer le conflit généré par la combinaison des données émanant de différentes sources imparfaites.

3. Décision : l'étape de combinaison produit un score pour chaque couple, dont plusieurs peuvent être en conflit. Les conflits sont résolus selon deux approches :

   - En choisissant le couple qui a le meilleur score. Par exemple, dans [30], après combinaison des masses, les couples qui se correspondent sont ceux avec la meilleure probabilité.
   - En utilisant une fonction d'optimisation. Dans [35], ce problème est transformé en recherche de cliques maximales dans un graphe. Dans [43], le problème est modélisé sous forme de problèmes de communication : émetteur, canal et récepteur.

2.3 Théorie de Dempster-Shafer

L'appariement d'objets se base sur la combinaison des mesures de similarités, qui sont calculées à partir des distances mesurées entre les objets de deux ou plusieurs sources. Or, pour les réseaux d'assainissement, les sources sont souvent imparfaites. L'appariement des objets à partir de plusieurs sources est un problème de prise de décision multicritères où les critères sont les différentes mesures de similarités associées aux différents candidats possibles. La théorie de DS permet de fusionner des données émanant de plusieurs sources. Aussi, elle est utilisée dans le domaine de prise de décisions multicritères. De plus, elle permet de prendre en considération l'incertitude, l'imprécision et l'incomplétude des données. Pour ces raisons, nous avons choisi d'utiliser la théorie de DS pour apparier les objets des réseaux d'assainissement. Nous allons présenter dans ce qui suit, les bases de cette théorie.

Fonction de masses

Soit $\Omega = \{H_1, H_2 \dots H_N\}$, l'ensemble de discernement, c'est-à-dire les choix possibles que nous pouvons faire en supposant un « monde fermé ». La vraisemblance associée à un sous-ensemble de Ω est définie par la fonction de masse $m(.)$:

$$m : 2^{\Omega} \mapsto [0, 1] \text{ avec } \sum_{A \subseteq \Omega} m(A) = 1 \tag{2.1}$$

A est appelé l'élément focal lorsque $m(A) > 0$. Les valeurs de $m(.)$ modélisent l'incertitude. De plus, comme A est un sous-ensemble de Ω, $m(A)$ représente l'imprécision lorsque la cardinalité de A est supérieure à 1. La valeur de masse d'un sous-ensemble $A = \{H_1, H_2, H_3\}$ n'est pas répartie de façon

équiprobable entre les singletons H_1, H_2 et H_3, contrairement à la théorie des probabilités. Ainsi, la valeur associée à $m(\Omega)$ représente l'ignorance et non une équiprobabilité.

Fonction de croyance et fonction de plausibilité

La fonction de croyance Cr(.) représente la vraisemblance minimale attachée à un sous ensemble A de Ω, telle que :

$$\mathrm{Cr}(\emptyset) = 0,\ \mathrm{Cr}(\Omega) = 1 \text{ et } \mathrm{Cr}(A) = \sum_{B \subseteq A} m(B) \tag{2.2}$$

La fonction de Plausibilité Pl(.) représente la vraisemblance maximale attachée à un sous ensemble A de Ω, telle que :

$$\mathrm{Pl}(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{2.3}$$

Affaiblissement

L'affaiblissement permet de prendre en considération le doute sur la fiabilité des informations qu'une source fournit. Pour un degré de fiabilité $r \in [0, 1]$ un taux d'affaiblissement de valeur $\alpha = 1 - r$, modifie $m(.)$ comme suit :

$$m^{\alpha}(A) = (1 - \alpha)m(A) \tag{2.4a}$$

$$m^{\alpha}(\Omega) = (1 - \alpha)m(\Omega) + \alpha \tag{2.4b}$$

Combinaison

Soit deux ou plusieurs sources d'informations qui s'expriment dans le même cadre de discernement Ω et communiquent des informations sur le même objet. Pour combiner ces masses en une seule, plusieurs opérateurs de combinaison sont définis. L'opérateur le plus populaire est celui de la conjonction. Soit $m_1(.)$ et $m_2(.)$ deux fonctions de masses qui représentent les vraisemblances associées à deux sources indépendantes et fiables, l'opérateur de conjonction est défini comme suit :

$$(m_1 \cap m_2)(A) = \sum_{B \cap C = A} m_1(B) \times m_2(C), \quad A \subseteq \Omega \tag{2.5}$$

L'opérateur de conjonction normalisé est défini par :

$$m_1 \oplus m_2 = \begin{cases} \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) \times m_2(C) & \text{si } A \neq \emptyset \\ 0 & \text{si } A = \emptyset \end{cases} \tag{2.6}$$

K est la masse associée au conflit entre les deux masses m_1 et m_2, définie par :

$$K = m(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{2.7}$$

Décision

La théorie des croyances propose plusieurs règles de décision, telles que la sélection de l'hypothèse avec la valeur maximale de croyance, de plausibilité ou de probabilité pignistique. La probabilité pignistique est définie dans [44] comme suit :

$$\mathrm{BetP}(H_i) = \sum_{A \subseteq \Omega, H_i \in A} \frac{m(A)}{|A|} \tag{2.8}$$

avec $|A|$ est le cardinal de A.

2.4 Contribution

Pour apparier les objets spatiaux des réseaux d'assainissement, émanant de plusieurs sources, nous avons proposé un processus basé sur le concept de stroke et la théorie de DS. Le concept de stroke est utilisé pour capturer la structure globale des branches, par conséquent réduire l'impact des nœuds manquants sur les résultats de l'appariement. La théorie DS est utilisée pour exécuter les opérations de combinaison en prenant en considération les imperfections des sources. Les mesures de similarité et les étapes de notre approche sont présentées dans les paragraphes suivants.

2.4.1 Les mesures de similarités

Afin d'évaluer la similarité entre les objets de différentes sources, nous proposons d'utiliser quatre distances de types géographiques et topologiques. Une mesure de similarité est dérivée de chaque distance.

- La distance de Hausdorff [45] : permet de mesurer la distance entre deux sous-ensembles d'un espace métrique, dans notre cas des lignes. Le choix de cette mesure est basé sur l'hypothèse que deux objets qui se correspondent sont proches. Elle est calculée comme suit :

$$d_{\mathrm{Hausdorff}}(L_1, L_2) = \max\{\min_{a \in L_1, b \in L_2} d_{\mathrm{Euclidean}}(a, b), \min_{a \in L_1, b \in L_2} d_{\mathrm{Euclidean}}(b, a)\}. \tag{2.9}$$

- La longueur : le choix de cette distance est basé sur l'hypothèse que deux objets qui se correspondent ont une taille similaire. Elle est calculée comme suit :

$$d_{\mathrm{length}}(L_1, L_2) = \frac{|\mathrm{length}(L_1) - \mathrm{length}(L_2)|}{\max(\mathrm{length}(L_1), \mathrm{length}(L_2))} \tag{2.10}$$

- L'orientation : deux objets qui se correspondent ont des orientations similaires. Elle est calculée comme l'angle entre chaque ligne et l'axe horizontal.

- Le degré des sommets : représente le nombre de lignes qui sont connectées à un nœud. Pour une ligne, nous considérons la moyenne des degrés des deux nœuds d'extrémité. Le choix de cette mesure est basé sur l'hypothèse que deux lignes qui se correspondent ont des voisins adjacents

similaires. Cette mesure est calculée comme suit :

$$d_{\text{degree}}(L_{a,b}, L_{c,d}) = \frac{d_{\text{degree}}(a,c) + d_{\text{degree}}(b,d)}{2}, \text{ Avec } \quad d_{\text{E}}(a,c) <= d_{\text{E}}(a,d) \tag{2.11}$$

et

$$d_{\text{degree}}(n_1, n_2) = \frac{|\text{degree}(n_1) - \text{degree}(n_2)|}{\max(\text{degree}(n_1), \text{degree}(n_2))} \tag{2.12}$$

d_{E} est la distance euclidienne.

Nous avons divisé ces mesures en deux catégories. La première inclut la distance de Hausdorff, la longueur et l'orientation. Dans cette catégorie quand les mesures ont des valeurs proches de 1, elles indiquent la non-correspondance entre les couples. Par contre quand leurs valeurs sont faibles elles ne permettent pas de confirmer la correspondance des objets car plusieurs candidats peuvent avoir des valeurs proches. La deuxième catégorie inclut seulement le degré des sommets. À l'opposé de la première catégorie, quand la mesure basée sur le degré des sommets est élevée la non-correspondance ne peut pas être confirmée, car elle peut être due aux données manquantes. Par contre quand cette valeur est faible cela indique que la vraisemblance que les objets se correspondent est élevée, car avoir plusieurs objets avec un voisinage similaire n'est pas fréquent, surtout dans le cas où les strokes sont utilisés comme unité de correspondance. Cette distinction sera considérée dans le processus de DS d'appariement.

2.4.2 Approche proposée pour l'appariement des objets spatiaux des réseaux d'assainissement

Processus d'appariement

Notre processus d'appariement des réseaux d'assainissement est illustré dans la Figure 2.1. La première étape consiste à construire les strokes à partir des données disponibles. Deux lignes sont considérées comme un même stroke si l'angle de déviation entre les deux ne dépasse pas 20 degrés.

La deuxième étape consiste à utiliser les strokes comme unité d'appariement pour surpasser le problème de nœuds manquants. Pour chaque stroke de la source de référence, une sélection des candidats dans la source cible est effectuée, en se basant sur une zone tampon ainsi que des filtres. Nous avons choisi une zone tampon de taille de 15 mètres de large, une valeur assez élevée pour éviter d'ignorer des potentielles correspondances. Ensuite, nous avons appliqué les filtres suivants :

- $d_{\text{Hausdorff}} \leq 40$ m
- $|\theta_{L_1} - \theta_{L_2}| \leq 45^{\text{o}}$
- $d_{\text{longueur}} \geq 0.8$ m

Les similarités entre les strokes sont évaluées en se basant sur la théorie de DS dans la troisième étape. Les détails de ce processus sont décrits dans le prochain paragraphe.

La quatrième et la cinquième étapes consistent à apparier les lignes qui appartiennent à des strokes qui n'ont pas eu de correspondants dans l'étape précédente. Comme pour les strokes, une zone tampon

et des filtres sont d'abord appliqués pour réduire les candidats, ensuite les opérations de combinaison et de décision sont faites à travers la théorie de DS.

Puisque les réseaux d'assainissements subissent fréquemment des opérations d'extension et de remplacement, les lignes qui n'ont pas de correspondants comprennent non seulement des conduites qui n'ont pas de correspondants en réalité, mais aussi des lignes qui ont une fausse représentation ou qui ont été installées récemment. C'est pourquoi, un appariement partiel est effectué en ajoutant uniformément des points fictifs provisoires.

Processus d'appariement basé sur la théorie de DS

La théorie de DS est utilisée pour combiner les mesures de similarités dans les troisième et cinquième étapes de notre processus. Nous proposons ici une amélioration du processus introduit dans [46] et utilisé pour l'appariement dans [31, 29]. Soit deux sources indépendantes comprenant respectivement N et M objets, nous définissons l'ensemble de discernement pour la combinaison locale des mesures comme $\Omega_1 = \{C_{i,j}, \neg C_{i,j}, E\}$, où $C_{i,j}$ est la masse associée au couple potentiel indexé par $i \in \{1 \dots N\}$ et $j \in \{1 \dots M\}$. $E = \{C_{1,1} \dots C_{N,M}\}$ est l'ensemble de discernement pour l'étape de combinaison des candidats.

Figure 2.2, illustre nos trois contributions principales.

1. Classement des candidats : nous avons modifié la première étape de transformation des distances en mesures, dans laquelle nous avons pris en compte le classement des candidats. Puisque les objets de références peuvent avoir plusieurs candidats très proches, il est important de classer ces derniers en fonction de chaque distance. Cette information est pertinente pour favoriser les candidats les plus proches. Nous proposons de classer les candidats en utilisant l'équation 2.13, inspiré du travail [34].

$$\text{Similarity}_{\text{metric}}(C_{i,j}) = \frac{d_{\text{metric}}(C_{i,j})^{-\beta}}{\sum_{k=1}^{\text{N}_\text{c}} d_{\text{metric}}(C_{i,k})^{-\beta}} \tag{2.13}$$

2. Modèles mixtes : contrairement aux études précédentes qui ont utilisé un seul modèle pour toutes les distances, nous distinguons entre les deux catégories de distances que nous avons introduites dans la section 2.4.2, par deux modèles de transformation différents. Nous avons choisi d'utiliser pour la première catégorie de mesures, le modèle 1 suivant :

$$m_i = \begin{cases} m_i(\{\text{C}_\text{i}\}) = 0 \\ m_i(\overline{\{\text{C}_\text{i}\}}) = r_i(1 - \text{L}_\text{i}) \\ m_i(\Omega) = 1 - r_i(1 - \text{L}_\text{i}) \end{cases} \tag{2.14}$$

Pour la deuxième catégorie le modèle 2 suivant :

$$m_i = \begin{cases} m_i(\{\text{C}_\text{i}\}) = r_i(\text{L}_\text{i}) \\ m_i(\overline{\{\text{C}_\text{i}\}}) = r_i(1 - \text{L}_\text{i}) \\ m_i(\Omega) = 1 - r_i \end{cases} \tag{2.15}$$

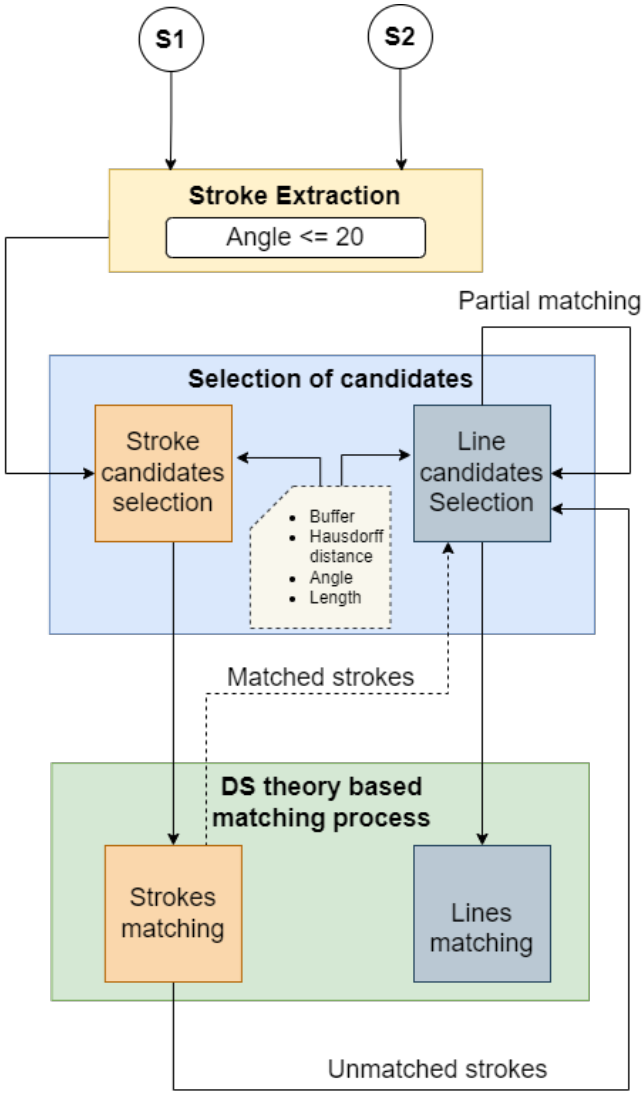


FIGURE 2.1 : Notre processus d'appariement des conduites des réseaux d'assainissement.

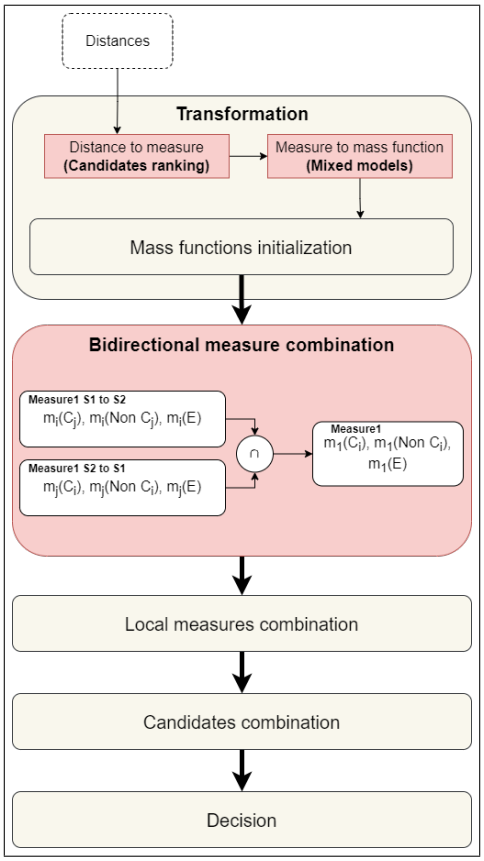
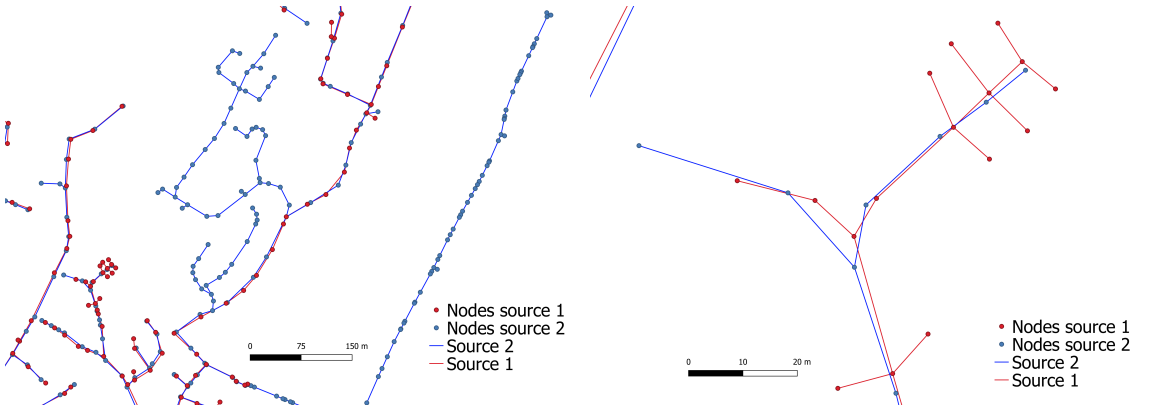


FIGURE 2.2 : Notre amélioration du processus d'appariement basé sur la théorie de DS.

r_i est le taux d'affaiblissement associé à C_i. L_i est la similarité calculée à partir de l'équation 2.13.

3. Combinaison bidirectionelle : nous considérons le classement des candidats lors de la première transformation des distances en mesures. Cependant, cet ordre peut varier selon la direction de l'appariement (référence/cible). Par conséquent, contrairement aux études précédentes où la combinaison locale est effectuée après l'initialisation des masses, nous proposons de combiner d'abord les masses dans les deux directions.

La combinaison locale et la combinaison des candidats sont effectuées comme dans [31, 29], où les masses sont combinées avec l'opérateur de conjonction normalisé. Ainsi, la décision est prise en fonction de la plausibilité maximale.



(a) Exemple de conduites manquantes dans la source1. (b) Exemple de conduites manquantes dans la source2.

FIGURE 2.3 : Exemples d'imperfections dans la base de données de Prades-Le-Lez.

2.5 Expérience et résultats

2.5.1 Données

Nous avons évalué chaque proposition du processus de combinaison basé sur la théorie de DS : classement des candidats, modèles mixte et combinaison bidirectionelle, en utilisant des exemples synthétiques. Ainsi, nous les avons comparé au travail publié dans [31]. Les détails de cette comparaison sont décrits dans le chapitre 3 du manuscrit.

Pour valider notre processus d'appariement dans sa globalité, nous avons utilisé deux bases de données réelles du réseau d'assainissement de Prades-Le-Lez, une petite ville d'environ 6000 habitants située dans le sud de la France. Les bases de données proviennent de deux organismes différents et ont été produites à des dates différentes : 2014 et 2017. Elles nous ont été fournies par les gestionnaires du réseau et elles contiennent respectivement 804 et 883 conduites représentant 23Km et 25.5Km de canalisations.

En premier lieu des écarts de représentation peuvent être observés dans Figure 2.3. De plus, malgré le fait que le réseau de 2017 soit plus récent et contient plus de conduites, nous pouvons remarquer un manque de données dans les deux sources (Figure 2.3a et Figure 2.3b). Ainsi, les nœuds qui représentent les positions des jonctions entre les conduites sont plus présents dans la source de 2017 que dans celle de 2014. Nous comparons nos résultats d'appariement à ceux réalisés manuellement par l'opérateur. Nous utilisons la précision et le rappel pour l'évaluation, définis comme suit :

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}} \tag{2.16}$$

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}} \tag{2.17}$$

2.5.2 Résultats

Nous avons suivi notre processus décrit dans Figure 2.1, pour trouver automatiquement les objets correspondants à partir des deux sources de données sur la ville de Prades-Le-Lez. Nous désignons respectivement les bases de données de 2014 et 2017 par source 1 et source 2. La première étape nous a permis de construire 398 et 421 strokes respectivement à partir de la source 1 et la source 2. Après sélection des candidats dans les deux sens, notre processus de combinaison basé sur la théorie de DS (Figure 2.2) a été effectué. En raison de l'imperfection des sources, nous avons combiné les quatre mesures de similarité avec une fiabilité de 0,8 pour chacune, laissant ainsi place à l'ignorance. Dans cette phase, 277 strokes ont été appariés. Ils représentent 576 canalisations (71,6%) de la source 1 et 580 canalisations (65,7%) de la source 2. Les strokes faussement appariés sont au nombre de 14, ce qui indique une précision d'appariement de 94,95%.

Le même processus a été appliqué directement aux conduites qui n'avaient pas de correspondants dans l'étape précédente. Compte tenu des nœuds manquants, nous avons réduit la fiabilité de la mesure du degré des sommets de 0,8 à 0,7. Dans cette étape 71 conduites de chaque source sont appariées. Ce nombre relativement faible est dû aux nœuds manquants qui affectent la longueur des lignes.

Pour résoudre la contrainte des nœuds manquants, nous avons effectué l'appariement partiel. Nous avons ajouté des nœuds fictifs tous les 16 mètres. Une valeur plus grande conduira à un nombre restreint de couples appariés et une valeur plus petite générera un grand nombre de sous-ensemble lors des étapes de combinaison.

Pour évaluer la qualité de l'appariement, les résultats ont été comparés à l'appariement effectué manuellement par les responsables du réseau. 731 conduites de la source 1 devraient avoir un correspondant dans la source 2. En utilisant notre proposition, nous avons trouvé 685 correspondances entières ou partielles entre les deux sources. Nous avons compté 16 faux positifs et 57 faux négatifs, soit une valeur de précision de 97,7% et un rappel de 92,3%. Pour garder une trace de l'incertitude liée aux couples appariés, la probabilité pignistique et les valeurs de plausibilités ont été conservées pour chaque appariement.

2.6 Conclusion

Dans ce chapitre nous avons présenté un nouveau processus pour l'appariement des objets des réseaux d'assainissement. Nous avons utilisé le concept de stroke et l'appariement partiel pour réduire l'impact des nœuds manquants sur les résultats de l'appariement. Nous avons utilisé quatre mesures de similarités pour évaluer la correspondance (distance de Hausdorff, longueur, orientation et degré des sommets). La combinaison des mesures a été conduite en se basant sur la théorie de Dempster-Shafer. Cette dernière a été choisie pour sa capacité à modéliser l'incertitude et l'imprécision des sources ainsi que pour les divers outils de combinaison qu'elle offre. Nous avons proposé trois modifications importantes du processus de DS pour l'appariement : classement des candidats, des modèles mixtes et appariement bidirectionel. Nous avons évalué nos choix sur des exemples synthétiques (voir le manuscrit de thèse) et réels. Les résultats montrent que notre processus fournit un bon score d'appariement avec une précision de 97% sur des données réelles.

Chapitre 3

Imputation des données des réseaux d'assainissement

3.1 Introduction

Un problème souvent rencontré lors de la gestion des systèmes d'information environnementaux, comme les réseaux souterrains, est celui des données manquantes [47, 48, 49, 50]. Les données manquantes rendent les interventions d'installation et l'anticipation des réparations difficiles, ce qui a un impact négatif sur la gestion des réseaux. Les gestionnaires prennent leur décision à partir des bases de données disponibles, qui souffrent généralement d'incomplétude, ce qui induit souvent à des retards des travaux publics, des embouteillages ou des dommages collatéraux lors des interventions. De plus, les experts en hydraulique qui étudient l'impact de certaines variables externes sur les réseaux, telles que le taux de rejet des abonnés dans le réseau, utilisent des logiciels de modélisation hydraulique qui nécessitent des bases de données complètes pour fonctionner avec succès.

Pour compléter ces données un nombre restreint d'études a été proposé dans la littérature, principalement basé sur des radars. Nous rappelons par exemple que dans [2] les auteurs cartographient les réseaux souterrains en utilisant des techniques de fusion bayésienne pour combiner des hypothèses collectées à partir d'un radar à pénétration de sol (RPS), des données d'inspection et des données issues des bases de données des opérateurs. L'étude dans [51], utilise aussi un modèle bayésien pour intégrer des données extraites à partir des capteurs et les bases de données disponibles. Dans [9], plusieurs capteurs ont été fusionnés. Par ailleurs, dans [6] les auteurs entraînent un réseau de neurones profond pour identifier des regards d'égout à partir d'images à haute résolution. Bien que ces propositions offrent des solutions innovantes pour collecter les données, elles sont coûteuses et nécessitent des moyens économiques et humains importants, ce qui n'est pas à la portée de toutes les municipalités, surtout pour les petites villes et communes.

Même si les cartes obtenues par ces propositions sont en bon accord avec les réseaux réels en termes de topologie, elles ne peuvent pas être utilisées directement par un logiciel de modélisation hydraulique. En effet, aucune ou très peu d'information sont disponibles sur les attributs principaux d'un réseau : diamètres des conduites, matériau, pentes etc. Une solution est alors de recourir aux

techniques d'imputation ou d'estimation des données manquantes, qui consistent à remplacer les valeurs manquantes par des valeurs estimées pour obtenir des bases de données complètes. Par exemple, les auteurs dans [52], ont utilisé les techniques d'imputation pour estimer les valeurs manquantes des attributs : diamètre, âge, nombre de connections et nombre des valves. Ils ont principalement utilisé des outils statistiques tels que la distribution des attributs, la moyenne, la médiane, l'algorithme d'espérance-maximisation ou la matrice de covariance. Bien que les résultats aient été encourageants pour certaines méthodes, cette étude avait des limitations importantes : les données estimées sont toutes discrètes et les tests ont été effectués sur un pourcentage faible de données manquantes, précisément entre 12.73% et 2.19%.

D'autres études ont été réalisées pour estimer les données manquantes dans d'autres domaines. Leur performances varient en fonction du type des données cibles : discrètes, continues ou mixtes [53], le pourcentage des données manquantes [54] ou le domaine d'application [55]. Historiquement, l'imputation des données a été effectuée à l'aide des techniques statistiques comme la régression linéaire multiple, la régression logistique, forêt d'arbres de décision ou l'inférence bayésiene [56, 47, 57, 58, 59, 60]. Aujourd'hui, l'imputation bénéficie du développement rapide des modèles d'apprentissage automatique comme les K plus proches voisins, machine à vecteurs de support, les réseaux de neurones artificiels [61, 62, 63, 64, 65] et récemment les réseaux de neurones en graphes (Graph Neural Network) [66]. Ces derniers sont particulièrement intéressants pour l'imputation des données manquantes pour les réseaux d'assainissement puisqu'ils utilisent la structure des graphes pour prédire les valeurs. Dans ce qui suit nous allons présenter deux contributions (la troisième contribution est présentée dans le document principale) concernant la complétion des données manquantes pour les réseaux d'assainissement.

3.2 Complétion des données pour une simulation hydraulique

Le présent travail vise à compléter automatiquement les données des réseaux d'assainissement nécessaire pour réaliser une simulation hydraulique, en utilisant le peu de données disponibles et les connaissances métiers. Cette contribution est organisée comme suit : La section 3.2.1 décrit la méthode et les données utilisées. Les résultats sont présentés dans la section 3.2.2 et la section 3.2.3 conclu cette étude.

3.2.1 Matériel et méthode

Nous supposons qu'une carte d'un réseau d'assainissement est disponible à partir d'un gestionnaire ou issue d'une opération de cartographie. Dans notre cas nous utilisons la carte de Prades-Le-Lez produite dans [67]. Les informations minimum nécessaires pour exécuter une simulation hydraulique sont : positions des nœuds d'entrées, débits associés à chaque nœud, géométrie des conduites, pente et rugosité. Les positions des nœuds sont considérées disponibles à partir de la carte. La rugosité est supposée uniforme et prend la valeur associée au matériel le plus utilisé dans la ville cible. N'ayant pas un impact considérable sur les résultats, la géométrie est supposée circulaire pour toutes les conduites. Les attributs qui restent à estimer sont : le diamètre, la pente, et le débit.

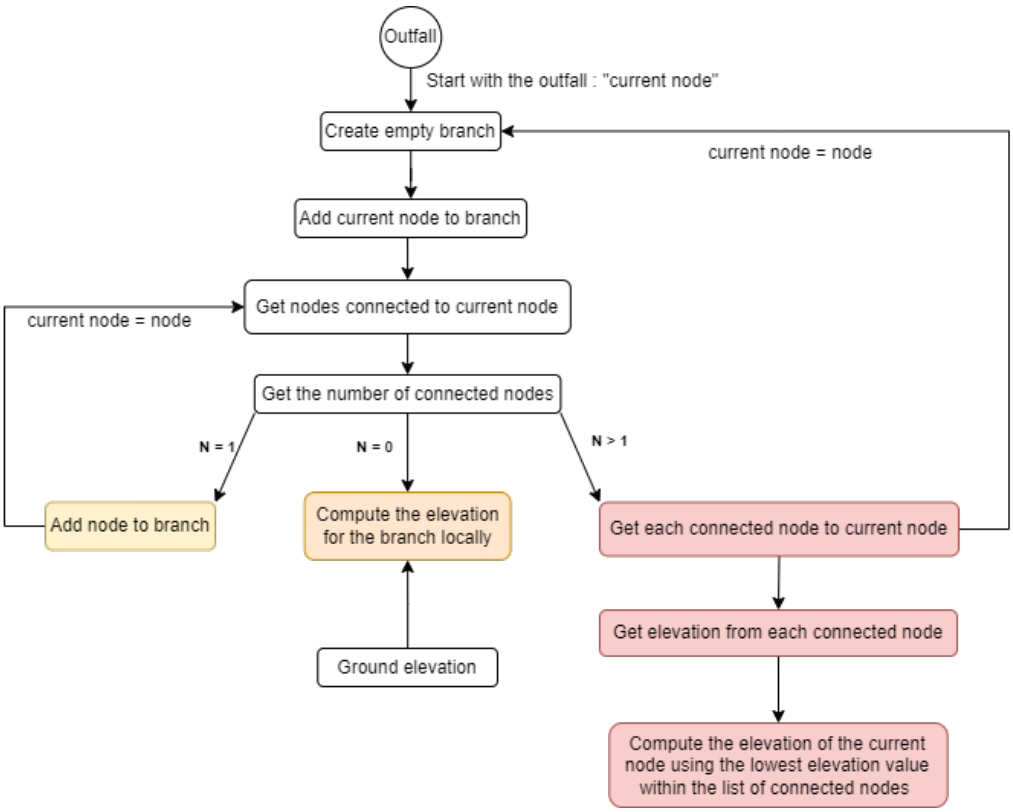


FIGURE 3.1 : Les étapes de l'estimation de la pente.

Estimation du diamètre. L'utilisateur indique au système une valeur maximale et minimale des diamètres. Ainsi, un algorithme associe à chaque conduite une valeur de diamètre dans cet intervalle selon l'ordre de Strahler [68]. Quand un minimum de 20% des valeurs des diamètres est disponible, l'apprentissage semi-supervisé décrit dans la Section 3.3 pourra être utilisé pour produire des meilleures estimations.

Estimation de la pente. L'algorithme développé attribue l'élévation du sol moins la profondeur minimale d'enterrement des canalisations aux nœuds, soit 0,8 m en France. Ensuite, puisque ces valeurs par défaut produisent des valeurs de pentes qui ne reflètent pas un écoulement gravitaire vers l'exutoire, la correction de ces valeurs est nécessaire. La figure 3.1, décrit les étapes de notre algorithme de correction des pentes.

Estimation du débit. Le débit associé à chaque nœud est estimé en fonction de la moyenne de consommation des citoyens des bâtiments proches du nœud. La moyenne de consommation a été estimée à partir de la consommation moyenne de l'eau potable. Ainsi, nous avons supposé que chaque bâtiment est probablement connecté à la conduite la plus proche. Les données des bâtiments sont généralement accessibles au public dans plusieurs pays.

Nous avons développé les outils nécessaires pour rendre toutes les opérations d'insertion des données

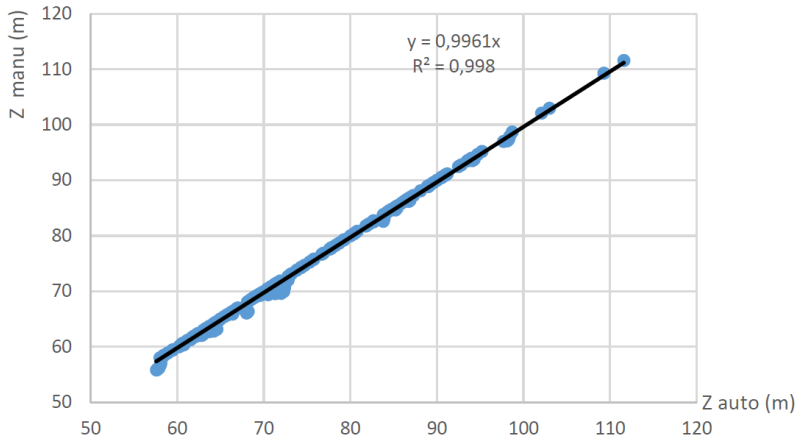


FIGURE 3.2 : Élévations automatiques (x-axis) et élévations manuelles (y-axis)

automatiques. Notre objectif est d'exécuter automatiquement une simulation hydraulique à partir des données que nous avons décrites. Nous avons utilisé le logiciel SWMM© [69] et nous avons appliqué notre approche sur Prades-Le-Lez.

3.2.2 Résultats et discussion

En se basant sur les algorithmes que nous avons décrits précédemment, nous avons réussi à compléter les données manquantes automatiquement et à exécuter les simulations hydrauliques sans erreur sur le logiciel SWMM© [69]. Comme nous n'avons pas de données de validation, les données estimées par notre proposition ont été aussi estimées manuellement par un ingénieur. Pour l'attribut pente qui est le plus difficile à acquérir, la figure 3.2, montre une corrélation proche de 1 entre les valeurs automatiquement estimées et celles manuelles. Les résultats des deux simulations par des données manuelles et automatiques montrent un comportement hydraulique cohérent (figure 3.3).

3.2.3 Conclusion et perspectives

Dans cette étude, nous avons proposé un processus pour estimer les données manquantes nécessaires pour exécuter une simulation hydraulique en utilisant une approche différente pour chaque paramètre requis. Nous avons basé nos algorithmes sur des informations souvent disponibles : l'élévation du sol, la moyenne des habitants par immeuble et la moyenne de consommation de l'eau potable journalière. En plus d'être totalement automatique, par conséquent rapide et facile à utiliser, les résultats ont produit un comportement hydraulique cohérent avec les données de validation manuelles.

Les données estimées ne sont pas exactes et ne reflètent pas systématiquement la réalité sur le terrain. Par conséquent, les résultats de la simulation sont inévitablement incertains et imprécis, et ils peuvent conduire à des conclusions erronées malgré un fonctionnement sans erreur. De plus, les approches d'estimation sont utilisées lorsque les données ne sont pas disponibles, et ils ne sont pas destinés à remplacer des données de terrain de bonne qualité. C'est pourquoi, nous travaillons sur

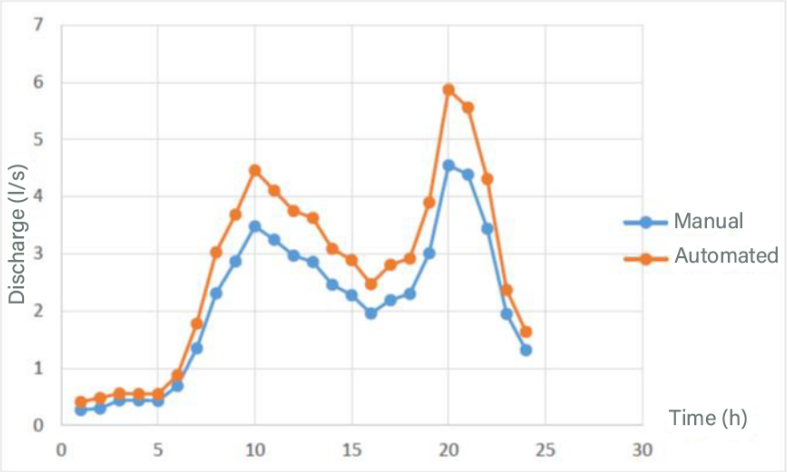


FIGURE 3.3 : Comparaison des hydrogrammes obtenus à partir de la méthode automatique et celle manuelle.

l'amélioration de cette proposition. Une des perspectives envisagées consiste à prendre en considération la portion des données disponibles lorsqu'elles existent, et compléter les données manquantes en se basant sur celles disponibles. La contribution suivante étudie cette problématique.

3.3 Imputation des données en utilisant les GNNs

L'objectif de cette contribution est d'utiliser les modèles de Graph Neural Network (GNN) pour compléter les données attributaires manquantes. Ce travail est organisé comme suit : la section 3.3.1 présente les modèles de Graph Neural Network. La section 3.3.2 décrit la méthode utilisée. Les résultats sont dans la section 3.3.3 et la conclusion en section 3.3.4.

3.3.1 Graph Neural Networks

Au cours de la dernière décennie, les modèles d'apprentissage automatique, en particulier les réseaux de neurones artificiels, ont été utilisés avec succès pour accomplir un ensemble de tâches complexes, comme le traitement automatique des langues [70], le traitement d'images [71] et la reconnaissance automatique de la parole [72]. Cependant, les modèles utilisés pour atteindre ce succès comme Convolutional Neural Network (CNN) ne sont adaptés qu'aux données euclidiennes et ne peuvent pas être appliqués directement aux graphes, puisque les structures des graphes ne sont pas fixes.

Plusieurs efforts ont été déployés pour que les graphes bénéficient de ce progrès et qu'il soit possible d'exploiter les structures des graphes dans le processus d'apprentissage. Historiquement, pour utiliser les graphes par des modèles d'apprentissage automatique, les utilisateurs encodent leur structures sous forme d'attributs à travers des méthodes statistiques [73, 74]. Cette approche est chronophage et dépend fortement du type d'application. C'est pourquoi les chercheurs ont proposé des alternatives automatiques et plus génériques.

Dans un premier temps, des approches nommées de Graph Embedding ont été développées pour transformer automatiquement les composants d'un graphe (nœuds/lignes) ou un graphe entier vers des vecteurs. L'objectif était de préserver dans ces vecteurs les relations entre les objets d'un graphe. Par exemple, deux nœuds qui se ressemblent dans un graphe doivent avoir des vecteurs d'embedding similaires, tandis que deux nœuds différents doivent avoir des vecteurs distincts. Les vecteurs sont ensuite utilisés par des algorithmes d'apprentissage. L'un des exemples le plus connu est l'algorithme Node2Vec [75] qui consiste à générer une représentation des nœuds en se basant sur des marches aléatoires(random walk). Récemment, pour palier les limites des méthodes d'embedding, principalement la perte d'information, les chercheurs ont proposé les Graph Neural Network qui sont des algorithmes d'apprentissage automatique qui opèrent directement sur les graphes. La première proposition était celle dans [76]. Ensuite plusieurs modèles, qui essayent de généraliser la notion de convolution dans les CNN vers les graphes, ont été publiés. Généralement, l'apprentissage se fait par transfert de messages entre les composants du graphe ; plus précisément, un nœud reçoit des messages à partir de ses voisins, ensuite ces messages sont agrégés pour construire automatiquement un vecteur d'embedding. Ce dernier sera exploité pour réaliser la tâche souhaitée : classification, régression, etc. Parmi les méthodes les plus populaires on trouve : GCN [77], ChebNet [78], GrapheSAGE [79] et TAGCN [80].

3.3.2 Matériel et Méthode

Dans ce travail, nous souhaitons compléter les valeurs d'attributs manquantes en s'appuyant sur les données disponibles et les structures des graphes, qui représentent les réseaux d'assainissement.

Les modèles et les configurations de test

Pour mettre en évidence la valeur ajoutée des GNN dans cette tâche de prédiction, nous allons comparer quatre modèles de GNN : GCN [77], ChebNet [78], GrapheSAGE [79] et TAGCN [80], à des algorithmes d'apprentissage populaires qui n'exploitent pas la structure des graphes dans la phase d'apprentissage : Machine à vecteurs de support (SVM) [81], les arbres de décisions (DT) [82], les réseaux de neurones multi-couches (ANN) [83].

Puisque les attributs diamètre et matériau sont nécessaires pour les interventions ainsi que les simulations hydrauliques, nous avons décidé de cibler ces deux attributs dans cette étude.

Les données manquantes dans les bases de données varient entre les fournisseurs, c'est pourquoi nous avons défini deux configurations, selon les données disponibles, pour mieux représenter les cas réels :

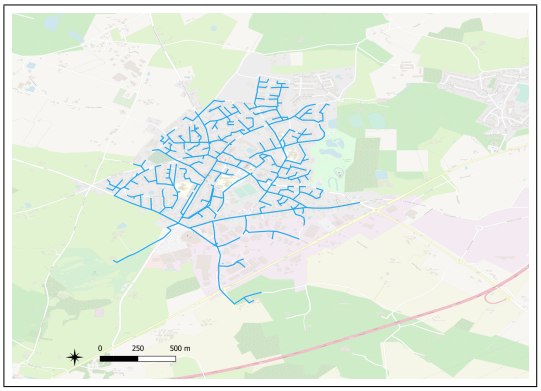
- Configuration 1 : le graphe du réseau d'assainissement, la partie disponible de l'attribut cible et l'attribut ordre de Strahler [68] associé à chaque conduite du réseau.
- Configuration 2 : le graphe du réseau d'assainissement, la partie disponible de l'attribut cible, l'attribut ordre de Strahler et d'autres attributs associés aux conduites.

Dans les deux configurations, les données sont divisées en deux ensembles : entraînement et test. L'ensemble d'entraînement contient les attributs qui sont à 100% renseignées pour chaque conduites ainsi que la portion disponible de l'attribut cible à compléter. Puisque les GNN prennent par nature

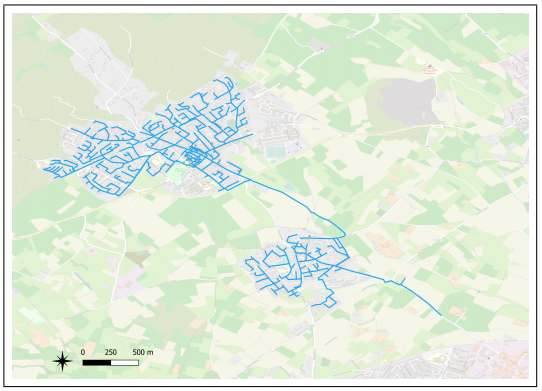
les structures des graphes, la matrice d'adjacence leur a été fournie comme entrée supplémentaire. 10% des données d'entraînement ont été utilisés pour la validation. Nous avons implémenté les réseaux de neurones multi-couches en utilisant Pytorch [84] avec 3 couches cachées. Nous avons utilisé les modèles machine à vecteurs support et les arbres de décision disponibles dans la bibliothèque Scikit-Learn [85]. Chacun des modèles GNN est composé de deux couches à convolution, et nous avons utilisé les modèles disponibles dans la bibliothèques Pytorch Géométrique [86].

Les données et la procédure des tests

Nous avons utilisé deux bases de données réelles pour comparer les modèles. La première est la base de données d'Angers métropole obtenue à partir des données publiques en France [5]. La deuxième est la base de données de Montpellier métropole issue des données publique de Montpellier [87]. Nous avons extrait à partir de ces deux sources deux réseaux qui contiennent tous les attributs nécessaires pour valider les prédictions. La Figure 3.4, montre les deux réseaux d'assainissement ; le réseau extrait d'Angers est composé de 754 conduites et le réseau extrait de Montpellier est composé de 1239 conduites. La Figure 3.5 indique la distribution des valeurs des attributs diamètre et matériau sur chaque réseau. Pour la configuration 2, nous avons utilisé l'attribut matériau dans la prédiction de l'attribut diamètre, et vice versa.



(a) Extrait du réseau d'assainissement de la métrople d'Angers [5].



(b) Extrait du réseau d'assainissement de la métropole de Montpellier [87].

FIGURE 3.4 : Les graphes utilisés dans cette étude.

Après les opérations de fine-tuning pour choisir les meilleurs paramètres des algorithmes d'apprentissages. Nous avons entraîné les modèles sur 90% des données et le test sur 10%. Pour observer et analyser le comportement des modèles lorsque les données d'entraînement varient, nous avons graduellement réduit de 10% l'ensemble d'entraînement en faveur de l'ensemble de test jusqu'à atteindre 10% d'entraînement et 90% de test. Les métriques adoptées pour évaluer la performance des modèles sont les suivantes :

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}} \tag{3.1}$$

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}} \tag{3.2}$$

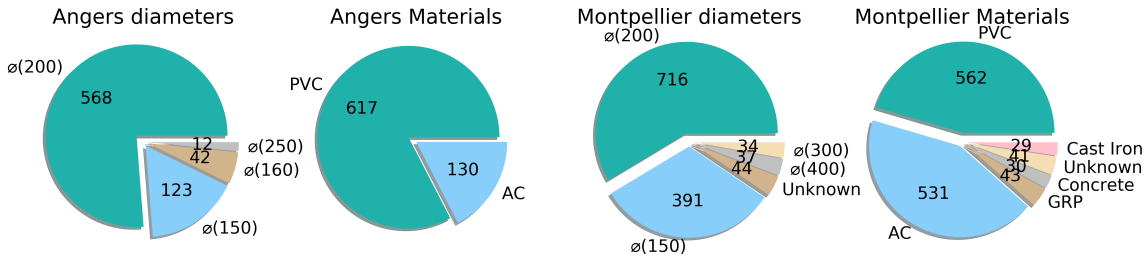


FIGURE 3.5 : Distribution des valeurs des attributs diamètre et matériau des extraits des bases de données d'Angers et de Montpellier (Seules les classes de plus de 10 éléments sont représentées).

$$\text{F1Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

$$\text{MacroF1Score} = \frac{1}{N} \sum_{i}^{N} \text{F1}_i \tag{3.4}$$

Pour chaque composition nous avons effectué 10 tests aléatoires et nous avons retenu comme scores les moyennes obtenues.

3.3.3 Résultats

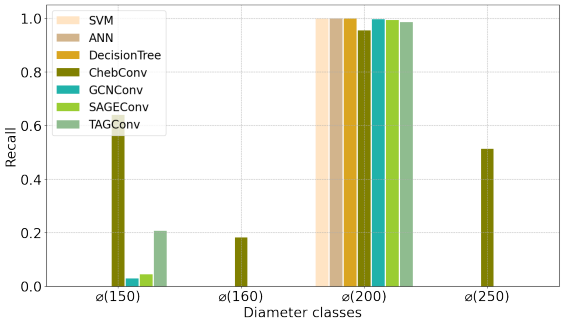
Dans cette section, nous allons montrer les résultats de la prédiction des attributs diamètre et matériau pour les deux configurations. Nous allons comparer les scores obtenus par les différents modèles SVM, DT, ANN, GCNConv, ChebConv, SAGEConv et TAGConv.

Pour la configuration 1, la Figure 3.6, montre les résultats sur la ville d'Angers. Nous remarquons que pour les deux attributs et sur les deux pourcentages affichés, les modèles de GNN donnent des meilleurs scores, particulièrement le modèle ChebConv qui a réussi à prédire toutes les classes même quand elles sont minoritaires. Des conclusions similaires peuvent être observées sur la Figure 3.7 pour la ville de Montpellier, où bien que GCNConv, SAGEConv et TAGConv ont des scores faibles pour les classes minoritaires, ils sont meilleurs par rapport aux modèles SVM, ANN et DT. Ainsi, ChebConv indique des scores importants pour toutes les classes.

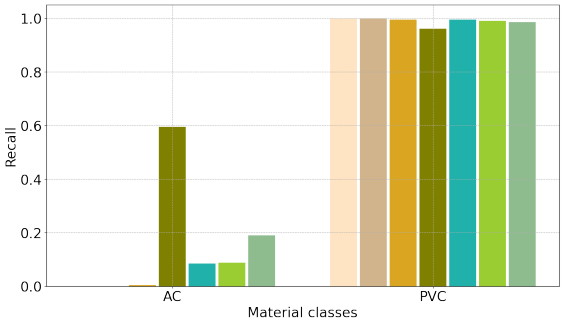
Il en est de même pour la configuration 2 où nous avons obtenu des meilleurs résultats en se basant sur les GNN. De plus, sur la Figure 3.8 nous remarquons que le modèle ChebConv permet de produire des scores élevés de prédiction même quand le taux des données manquantes est important. Par exemple, sur les données de Montpellier et avec un pourcentage de données manquantes de 60% le diamètre a été prédit avec un score Macro-F1 de 0.90%, et l'attribut matériau avec un Macro-F1 de 0.75%. Encore, les modèles GNN ont tendance à s'améliorer quand le taux des données disponibles augmente comparés aux autres modèles dont les performances restent presque en plateau.

3.3.4 Conclusion

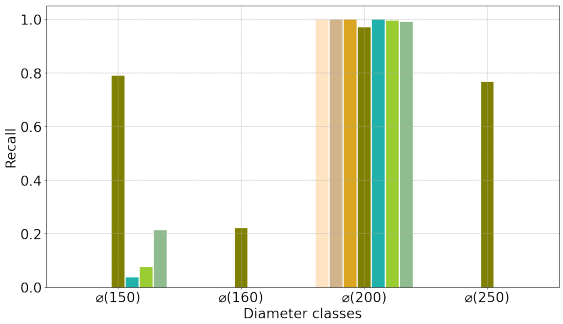
Dans cette étude nous avons enquêté sur l'efficacité des algorithmes d'apprentissage automatique basés sur des graphes dans le domaine de l'imputation des données manquantes. Nous avons comparé 3



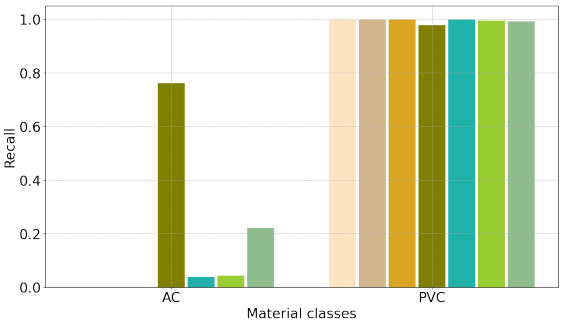
(a) Prédiction du diamètre : 30% entraînement and 70% test.



(b) Prédiction du matériau : 30% entraînement and 70% test.

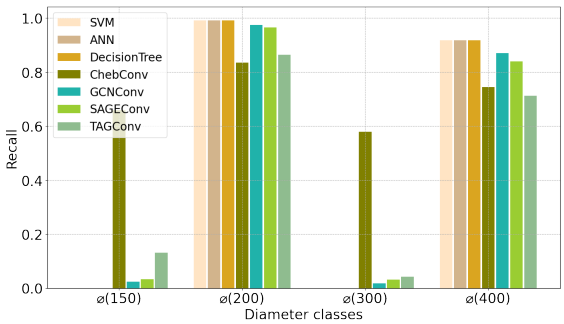


(c) Prédiction du diamètre : 70% entraînement and 30% test.

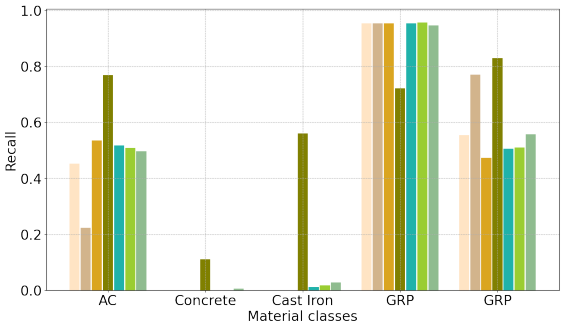


(d) Prédiction du matériau : 70% entraînement and 30% test.

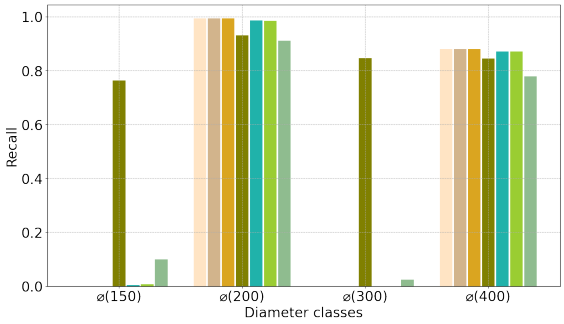
FIGURE 3.6 : Configuration 1 : prédiction des attributs diamètre et matériau pour la base de données d'Angers pour chaque classe, évaluation par la métrique : recall.



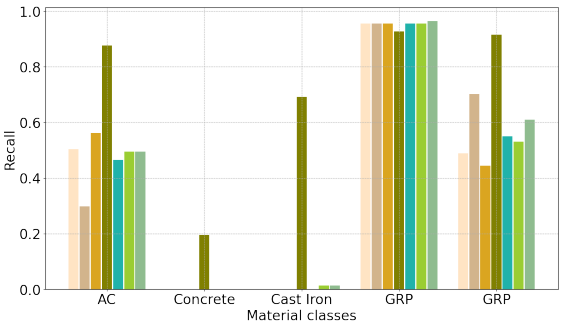
(a) Prédiction du diamètre : 30% entraînement and 70% test.



(b) Prédiction du matériau : 30% entraînement and 70% test.

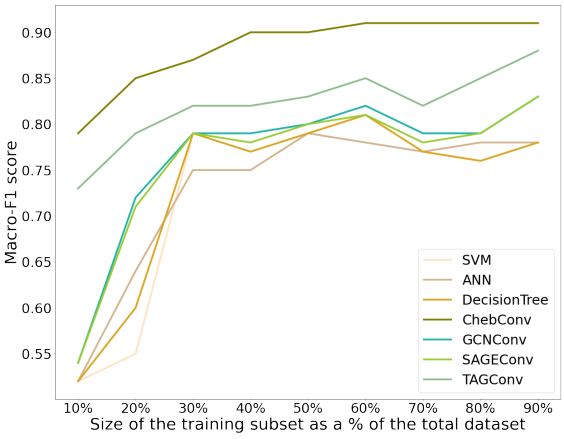


(c) Prédiction du diamètre : 70% entraînement and 30% test.

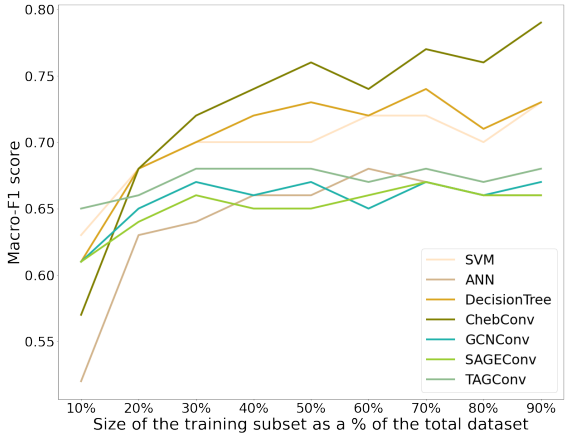


(d) Prédiction du matériau : 70% entraînement and 30% test.

FIGURE 3.7 : Configuration 1 : prédiction des attributs diamètre et matériau pour la base de données de Montpellier pour chaque classe, évaluation par la métrique : recall



(a) L'attribut diamètre.



(b) L'attribut matériau.

FIGURE 3.8 : L'évolution de la performance (score F1) des modèles en fonction de la réduction du taux des données manquantes pour la base de données de Montpellier sur la configuration 2.

modèles qui n'utilisent pas la structure des graphes dans leurs processus d'apprentissage : machine à vecteurs support, les arbres de décision et les réseaux de neurones multi-couches, avec des modèles qui exploitent par défaut la structure : GCN, ChebNet, GraphSage et TAGCN. Les résultats ont montré clairement la supériorité des GNN, particulièrement quand le pourcentage des données manquantes est important. ChebNet a pu obtenir les meilleurs résultats dans les deux configurations que nous avons définies. Nos tests étaient sur des données réelles de deux villes en France : Angers et Montpellier, ce qui démontre que cette proposition pourra aider les gestionnaires des réseaux d'assainissement à compléter les données manquantes dans leur base de données.

Conclusion

Pour assurer un service continu à la population, la gestion des réseaux souterrains, notamment ceux d'assainissement, implique une expertise dans différents domaines : économie, ressources humaines, gestion, ingénierie, développement de logiciels, etc. La revue de la littérature réalisée dans le cadre de ce projet de thèse a montré que les problèmes liés aux données font partie des défis quotidiens auxquels les opérateurs sont confrontés. Ils peuvent aussi causer des incidents, des embouteillages, des dépassements de budget, des risques pour l'environnement et les travailleurs, etc. Nous avons focalisé nos travaux sur les enjeux liés aux données. Nous les avons résumés en trois caractéristiques :

- Les données sont collectées à partir de plusieurs sources.
- Les données sont hétérogènes.
- Les données sont imparfaites : incertaines, imprécises et incomplètes.

Compte tenu de ces caractéristiques, nous avons établi que pour obtenir des bases de données des réseaux d'assainissement plus précises et complètes, des opérations de fusion de données et de complétion des données manquantes sont nécessaires. Après l'étude de ces deux domaines nous avons identifié 4 questions scientifiques :

- Dans la perspective de conduite des opérations de fusion de données, comment les sources de données des réseaux d'assainissement peuvent-elles être modélisées compte tenu de leur nature hétérogène et imparfaite ?
- Comment réaliser l'appariement d'objets des réseaux d'assainissement et comment modéliser les imperfections ?
- Comment estimer les données manquantes des réseaux d'assainissement, et comment la structure des réseaux peut-elle être exploitée à cette fin ?
- Quelles connaissances métiers pourraient être utilisées à la fois pour la fusion des données et l'estimation des données manquantes ?

Dans le **Chapite1**, nous avons abordé la première question en proposant un méta-modèle des sources de données des réseaux d'assainissement et sa relation avec les modèles métiers. Notre proposition était motivée par deux facteurs. Premièrement, l'hétérogénéité des sources de données : SIG, images, GPR etc, nécessite un cadre unifié pour faciliter les opérations de fusion de données. Deuxièmement, nous avons choisi un méta-modèle plutôt qu'un modèle simple car les sources de données peuvent évoluer dans le temps, ainsi une modélisation exhaustive des sources n'est pas générique. Notre proposition est inspirée du domaine du Big Data qui se caractérise aussi par l'hétérogénéité et la multitude des sources. Nous avons divisé les sources de données en 3 catégories : structurées, semi-structurées et non structurées. Nous avons implémenté ce meta-modèle sur la plate-forme Moose et nous avons montré en utilisant des exemples concrets que le méta-modèle est générique et peut être utilisé pour réaliser des opérations de fusion.

Dans le **Chapitre2**, nous avons abordé la deuxième et la quatrième question. Nous avons proposé un nouveau processus pour l'appariement des conduites des réseaux d'assainissement. Pour réduire

l'impact des données manquantes précisément des nœuds, dans un premier temps nous avons utilisé les strokes comme unité d'appariement, ensuite nous avons eu recours à un appariement partiel. La modélisation des imperfections a été conduite à travers la théorie de Dempster-Shafer. Une amélioration du processus d'appariement basé sur la théorie DS a été introduite. Ainsi, les résultats sur les données synthétiques et réelles ont montré que notre proposition peut aider les gestionnaires à fusionner leur données efficacement.

Dans le **Chapitre3**, nous avons répondu à la troisième et la quatrième question. Nous avons proposé des algorithmes pour compléter les données manquantes nécessaires pour réaliser une simulation hydraulique. Les résultats ont montré que malgré les imperfections de ces estimations, elles permettent aux spécialistes de réaliser des simulations, et d'analyser le comportement des réseaux d'assainissement. Ensuite, à travers plusieurs tests nous avons cherché à savoir si la structure des réseaux pourrait aider à l'estimation des données manquantes. Les résultats ont démontré que les Graph Neural Network, qui sont des algorithmes d'apprentissage automatique qui exploitent la structure des graphes dans le processus d'apprentissage, permettent de prédire efficacement les données manquantes comparés aux autres modèles qui n'exploitent pas cette structure.

Les résultats présentés dans ce travail de thèse ouvrent de nouvelles perspectives de recherche intéressantes. Ainsi notre méta-modèle a été testé à deux reprises ; la première quand nous avons instancié une source semi-structurée et deux sources non-structurées, la deuxième est lorsque nous avons effectué l'appariement des objets sur deux sources structurées. Cependant, compte tenu des données disponibles, nous n'avons pas pu l'évaluer sur d'autres sources non-structurées, en particulier celles basées sur des données radar. Ainsi, l'instanciation et l'implémentation des composants nécessaires sont à définir par les futurs utilisateurs et chercheurs. Pour l'appariement spatial, comme indiqué dans le COVADIS les canalisations et les nœuds peuvent être divisés en plusieurs types. Dans l'état actuel de notre proposition, cet aspect n'est pas considéré. Par conséquent, les futures études pourront être orientées dans cette direction.

Le troisième chapitre est celui dans lequel nous avons utilisé le plus les connaissances métier. Nous sommes certains que d'autres informations pourront être exploitées, comme le type de bâtiment proche d'une conduite pour inférer le diamètre, ou des rapports publics pour identifier le matériau. Un exemple concret de la collecte de connaissances sur les réseaux est celui dans [88], où les données sont extraites automatiquement à partir du Web. Néanmoins, la contribution la plus importante concernant l'imputation des données consistait à montrer le rôle de la structure des réseaux dans l'estimation des données manquantes. Pour la complétion des données attributaires nous nous sommes limités à des données discrètes. L'attribut pente – le seul attribut nécessaire pour la simulation hydraulique ayant des valeurs continues – n'était pas étudié. C'est pourquoi, les études futures peuvent être orientées vers l'estimation de la pente en se basant sur un modèle GNN de régression.

Références du résumé

[1] F. Pisani, "Voyage dans les villes intelligentes: entre datapolis et participolis," *Netexplo, Paris*, 2015.

[2] H. Chen and A. G. Cohn, "Buried utility pipeline mapping based on multiple spatial data sources: A bayesian data fusion approach," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[3] (2022) Ineris. https://www.reseaux-et-canalisations.ineris.fr/gu-presentation/construire-sans-detruire/prevenir-les-risques.html. Accessed 20 september 2022.

[4] W. McMahon, M. Burtwell, and M. Evans, "Minimising street works disruption: the real costs of street works to the utility industry and society. london: Uk water industry research," 2005.

[5] (2015) Open platform for french public data. https://www.data.gouv.fr/. Accessed 01 august 2020.

[6] B. Commandre, D. En-Nejjary, L. Pibre, M. Chaumont, C. Delenne, and N. Chahinian, "Manhole Cover Localization in Aerial Images with a Deep Learning Approach," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42W1, pp. 333–338, May 2017.

[7] C. Abdelbaki and M. Zerouali, "Modélisation d'un réseau d'assainissement et contribution a sa gestion a l'aide d'un système d'information géographique-Cas du chef lieu de commune de Chetouane-wilaya de Tlemcen Algérie," *LARHYSS Journal P-ISSN 1112-3680/E-ISSN 2521-9782*, no. 10, 2012.

[8] COVADIS. (2019) Standard de données réseaux d'AEP & d'assainissement, version 1.2. http://www.geoinformations.developpement-durable.gouv.fr/. Accessed 01 August 2020.

[9] M. Hafsi, P. Bolon, and R. Dapoigny, "Detection and localization of underground networks by fusion of electromagnetic signal and GPR images," in *Thirteenth International Conference on Quality Control by Artificial Vision 2017*, H. Nagahara, K. Umeda, and A. Yamashita, Eds., vol. 10338, International Society for Optics and Photonics. SPIE, 2017, pp. 7 – 14. [Online]. Available: https://doi.org/10.1117/12.2266946

[10] A.-C. Boury-Brisset, "Managing semantic big data for intelligence." in *Semantic Technologies for Intelligence, Defense, and Security*, 2013, pp. 41–47.

[11] X. L. Dong and D. Srivastava, "Big data integration," in *2013 IEEE 29th international conference on data engineering (ICDE)*. IEEE, 2013, pp. 1245–1248.

[12] A. Erraissi and A. Belangour, "Data sources and ingestion big data layers: meta-modeling of key concepts and features," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 3607–3612, 2018.

[13] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.

[14] Object Management Group. (2006) Meta object facility (MOF) 2.0 core specification. https://www.omg.org/spec/MOF/2.0/. Accessed 1 august 2020.

[15] A. R. Da Silva, "Model-driven engineering: A survey supported by the unified conceptual model," *Computer Languages, Systems & Structures*, vol. 43, pp. 139–155, 2015.

[16] J. M. Gascueña, E. Navarro, and A. Fernández-Caballero, "Model-driven engineering techniques for the development of multi-agent systems," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 1, pp. 159–173, 2012.

[17] A. Bertolino, A. Calabrò, F. Lonetti, A. Di Marco, and A. Sabetta, "Towards a model-driven infrastructure for runtime monitoring," in *Software Engineering for Resilient Systems*, E. A. Troubitsyna, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 130–144.

[18] T. B. la Fosse, Z. Cheng, J. Rocheteau, and J. M. Mottu, "Model-driven engineering of monitoring application for sensors and actuators networks," in *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2020, pp. 553–560.

[19] S. Ducasse, T. Gîrba, M. Lanza, and S. Demeyer, "Moose: A collaborative and extensible reengineering environment." in *Tools for Software Maintenance and Reengineering*. Citeseer, 2005.

[20] Moose. https://moosetechnology.org. Accessed 1 august 2020.

[21] Pharo. https://pharo.org/. Accessed 1 august 2020.

[22] Y. Belghaddar, A. Seriai, A. Begdouri, C. Delenne, N. Chahinian, R. Bachar, and M. Derras, "Towards a generic fusion framework for underground networks involving model-driven engineering domain," *International Journal of Information Science & Technology*, May 2022, <hal-03685139>. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03685139

[23] T. Devogele, C. Parent, and S. Spaccapietra, "On spatial database integration," *International Journal of Geographical Information Science*, vol. 12, no. 4, pp. 335–352, 1998.

[24] X. Tong, W. Shi, and S. Deng, "A probability-based multi-measure feature matching method in map conflation," *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5453–5472, 2009.

[25] J. O. Kim, K. Yu, J. Heo, and W. H. Lee, "A new method for matching objects in two different geospatial datasets based on the geographic context," *Computers & Geosciences*, vol. 36, no. 9, pp. 1115–1122, 2010.

[26] W. Song, J. M. Keller, T. L. Haithcoat, and C. H. Davis, "Relaxation-based point feature matching for vector map conflation," *Transactions in GIS*, vol. 15, no. 1, pp. 43–60, 2011.

[27] S. Volz, "An iterative approach for matching multiple representations of street data," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 36, no. Part 2/W40, pp. 101–110, 2006.

[28] A. Appriou, *Uncertainty theories and multisensor data fusion.* John Wiley & Sons, 2014.

[29] G. Nassreddine, F. Abdallah, and T. Denoeux, "Map matching algorithm using interval analysis and dempster-shafer theory," in *2009 IEEE Intelligent Vehicles Symposium.* IEEE, 2009, pp. 494–499.

[30] Y. Deng, A. Luo, J. Liu, and Y. Wang, "Point of interest matching between different geospatial datasets," *ISPRS International Journal of Geo-Information*, vol. 8, no. 10, p. 435, 2019.

[31] A.-M. O. Raimond, S. Mustière, and A. Ruas, "Knowledge formalization for vector data matching using belief theory," *J. Spatial Inf. Sci.*, vol. 10, pp. 21–46, 2015.

[32] B. Rosen and A. Saalfeld, "Match criteria for automatic alignment," in *Proceedings of 7th international symposium on computer-assisted cartography (Auto-Carto 7)*, 1985, pp. 1–20.

[33] L. Li and M. F. Goodchild, "An optimisation model for linear feature matching in geographical data conflation," *International Journal of Image and Data Fusion*, vol. 2, no. 4, pp. 309–328, 2011.

[34] C. Beeri, Y. Kanza, E. Safra, and Y. Sagiv, "Object fusion in geographic information systems," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 2004, pp. 816–827.

[35] A. Samal, S. Seth, and K. Cueto 1, "A feature-based approach to conflation of geospatial sources," *International Journal of Geographical Information Science*, vol. 18, no. 5, pp. 459–489, 2004.

[36] E. M. Xavier, F. J. Ariza-López, and M. A. Urena-Camara, "A survey of measures and methods for matching geospatial vector datasets," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–34, 2016.

[37] Y. Wang, D. Chen, Z. Zhao, F. Ren, and Q. Du, "A back-propagation neural network-based approach for multi-represented feature matching in update propagation," *Transactions in GIS*, vol. 19, no. 6, pp. 964–993, 2015.

[38] J. Wu, Y. Wan, Y.-Y. Chiang, Z. Fu, and M. Deng, "A matching algorithm based on voronoi diagram for multi-scale polygonal residential areas," *IEEE Access*, vol. 6, pp. 4904–4915, 2018.

[39] A. Saalfeld, "Conflation automated map compilation," *International Journal of Geographical Information System*, vol. 2, no. 3, pp. 217–228, 1988.

[40] B. Costes, "Matching Old Hydrographic Vector Data from Cassini's Maps," *e-Perimetron*, vol. 9, no. 2, pp. 51–65, 2014.

[41] M. Zhang, W. Shi, and L. 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*, vol. 9. Citeseer, 2005, pp. 101–110.

[42] S. Wang, Q. Guo, X. Xu, and Y. Xie, "A study on a matching algorithm for urban underground pipelines," *ISPRS International Journal of Geo-Information*, vol. 8, no. 8, p. 352, 2019.

[43] V. Walter and D. Fritsch, "Matching spatial data sets: a statistical approach," *International Journal of geographical information science*, vol. 13, no. 5, pp. 445–473, 1999.

[44] P. Smets and R. Kennes, "The transferable belief model," *Artificial intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[45] W. Rucklidge, *Efficient Visual Recognition Using the Hausdorff Distance*. Berlin, Heidelberg: Springer-Verlag, 1996.

[46] A. 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,*, pp. 27–40, 1991.

[47] P. Lin and X. X. Yuan, "A two-time-scale point process model of water main breaks for infrastructure asset management," *Water Research*, vol. 150, pp. 296–309, 2019. [Online]. Available: https://doi.org/10.1016/j.watres.2018.11.066

[48] D. T. Kofinas, A. Spyropoulou, and C. S. Laspidou, "A methodology for synthetic household water consumption data generation," *Environmental Modelling & Software*, vol. 100, pp. 48–66, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815216310520

[49] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*, vol. 38, no. 18, pp. 2895–2907, 2004.

[50] T. Schneider, "Analysis of incomplete climate data: Estimation of Mean Values and covariance matrices and imputation of Missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[51] M. Bilal, W. Khan, J. Muggleton, E. Rustighi, H. Jenks, S. R. Pennock, P. R. Atkins, and A. Cohn, "Inferring the most probable maps of underground utilities using bayesian mapping model," *Journal of Applied Geophysics*, vol. 150, pp. 52–66, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092698511730143X

[52] G. Kabir, S. Tesfamariam, J. Hemsing, and R. Sadiq, "Handling incomplete and missing data in water network database using imputation methods," *Sustainable and Resilient Infrastructure*, vol. 5, no. 6, pp. 365–377, 2020.

[53] C.-F. Tsai and F.-Y. Chang, "Combining instance selection for better missing value imputation," *Journal of Systems and Software*, vol. 122, pp. 63–71, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121216301583

[54] A. W.-C. Liew, N.-F. Law, and H. Yan, "Missing value imputation for gene expression data: computational techniques to recover missing data from available information," *Briefings in Bioinformatics*, vol. 12, no. 5, pp. 498–513, 12 2010. [Online]. Available: https://doi.org/10.1093/bib/bbq080

[55] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, Mar. 2010. [Online]. Available: https://doi.org/10.1007/s00521-009-0295-6

[56] R. H. Ngouna, R. Ratolojanahary, K. Medjaher, F. Dauriac, M. Sebilo, and J. Junca-Bourié, "A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values," *Engineering Applications of Artificial Intelligence*, vol. 95, no. July, p. 103822, 2020. [Online]. Available: https://doi.org/10.1016/j.engappai.2020.103822

[57] S. Bischof, A. Harth, B. Kämpgen, A. Polleres, and P. Schneider, "Enriching integrated statistical open city data by combining equational knowledge and missing value imputation," *Journal of Web Semantics*, vol. 48, pp. 22–47, 2018. [Online]. Available: https://doi.org/10.1016/j.websem.2017.09.003

[58] M. L. Yadav and B. Roychoudhury, "Handling missing values: A study of popular imputation packages in R," *Knowledge-Based Systems*, vol. 160, no. April, pp. 104–118, 2018. [Online]. Available: https://doi.org/10.1016/j.knosys.2018.06.012

[59] R. Serrano-Notivoli, M. de Luis, and S. Beguería, "An R package for daily precipitation climate series reconstruction," *Environmental Modelling & Software*, vol. 89, pp. 190–195, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136481521630281X

[60] M. Murtojärvi, T. Suominen, E. Uusipaikka, and O. S. Nevalainen, "Optimising an observational water monitoring network for Archipelago Sea, South West Finland," *Computers and Geosciences*, vol. 37, no. 7, pp. 844–854, 2011.

[61] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," *Neurocomputing*, vol. 72, no. 7, pp. 1483–1493, 2009, advances in Machine Learning and Computational Intelligence. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231209000149

[62] S. Belda, L. Pipia, P. Morcillo-Pallarés, J. P. Rivera-Caicedo, E. Amin, C. De Grave, and J. Verrelst, "DATimeS: A machine learning time series GUI toolbox for gap-filling and vegetation phenology trends detection," *Environmental Modelling & Software*, vol. 127, p. 104666, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815219310680

[63] J. Ma, J. C. Cheng, Y. Ding, C. Lin, F. Jiang, M. Wang, and C. Zhai, "Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series," *Advanced Engineering Informatics*, vol. 44, no. March, p. 101092, 2020. [Online]. Available: https://doi.org/10.1016/j.aei.2020.101092

[64] L. Giustarini, O. Parisot, M. Ghoniem, R. Hostache, I. Trebs, and B. Otjacques, "A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records," *Environmental Modelling & Software*, vol. 82, pp. 308–320, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.envsoft.2016.04.013

[65] F. V. Nelwamondo, D. Golding, and T. Marwala, "A dynamic programming approach to missing data estimation using neural networks," *Information Sciences*, vol. 237, pp. 49–58, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2009.10.008

[66] I. Spinelli, S. Scardapane, and A. Uncini, "Missing data imputation with adversarially-trained graph convolutional networks," *Neural Networks*, vol. 129, pp. 249–260, 2020. [Online]. Available: https://doi.org/10.1016/j.neunet.2020.06.005

[67] N. Chahinian, C. Delenne, B. Commandre, M. Derras, L. Deruelle, and J.-S. Bailly, "Automatic mapping of urban wastewater networks based on manhole cover locations," *Computers, Environment and Urban Systems*, vol. 78, p. 101370, 2019.

[68] A. Strahler, "Quantitative analysis of watershed geomorphology," *Eos, Transactions American Geophysical Union*, vol. 38, no. 6, pp. 913–920, 1957. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/TR038i006p00913

[69] Environmental Protection Agency, Storm Water Management Model. Epaswmm. https://www.epa.gov/. Accessed 1 august 2020.

[70] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 160–167. [Online]. Available: https://doi.org/10.1145/1390156.1390177

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[72] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[73] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[74] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," 2019.

[75] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.

[76] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[77] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017.

[78] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," 2017.

[79] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in neural information processing systems*, 2017, pp. 1024–1034.

[80] J. Du, S. Zhang, G. Wu, J. M. Moura, and S. Kar, "Topology adaptive graph convolutional networks," *arXiv preprint arXiv:1710.10370*, 2017.

[81] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[82] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[83] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[84] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS 2017 Workshop on Autodiff*, 2017. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ

[85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[86] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," *arXiv preprint arXiv:1903.02428*, 2019.

[87] Montpellier Méditeranée Métropole. Open data. https://data.montpellier3m.fr/. Accessed 1 august 2020.

[88] N. Chahinian, T. Bonnabaud La Bruyère, F. Frontini, C. Delenne, M. Julien, R. Panckhurst, M. Roche, L. Sautot, L. Deruelle, and M. Teisseire, "Weir-p: An information extraction pipeline for the wastewater domain," in *International Conference on Research Challenges in Information Science*. Springer, 2021, pp. 171–188.

Bibliography

Abdelbaki, C., & Zerouali, M. (2012). Modélisation d'un réseau d'assainissement et contribution a sa gestion a l'aide d'un système d'information géographique-Cas du chef lieu de commune de Chetouane-wilaya de Tlemcen Algérie. *LARHYSS Journal P-ISSN 1112-3680/E-ISSN 2521-9782*, (10).

Aissia, M.-A. B., Chebana, F., & Ouarda, T. B. (2017). Multivariate missing data in hydrology–review and applications. *Advances in Water Resources*, *110*, 299–309.

Alabadla, M., Sidi, F., Ishak, I., Ibrahim, H., Affendey, L. S., Ani, Z. C., Jabar, M. A., Bukar, U. A., Devaraj, N. K., Muda, A. S., et al. (2022). Systematic review of using machine learning in imputing missing values. *IEEE Access*.

Al-Bayati, A. J., & Panzer, L. (2019). Reducing damage to underground utilities: Lessons learned from damage data and excavators in north carolina. *Journal of Construction Engineering and Management*, *145*(12), 04019078.

Alt, H., Knauer, C., & Wenk, C. (2004). Comparison of distance measures for planar curves. *Algorithmica*, *38*(1), 45–58. https://doi.org/doi:10.1007/s00453-003-1042-5

American Society of Civil Engineers. (2021). A comprehensive assessment of america's infrastructure. *ASCE*.

Appriou, A. (1991). 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,* 27–40.

Appriou, A. (1998). *Uncertain data aggregation in classification and tracking processes*. Springer. https://doi.org/doi:10.1007/978-3-7908-1889-5_13

Appriou, A. (2014). *Uncertainty theories and multisensor data fusion*. John Wiley & Sons.

Armina, R., Zain, A. M., Ali, N. A., & Sallehuddin, R. (2017). A review on missing value estimation using imputation algorithm. *Journal of Physics: Conference Series*, *892*(1), 012004.

Assi, A., & Dhifli, W. (2021). Instance matching in knowledge graphs through random walks and semantics. *Future Generation Computer Systems*, *123*, 73–84. https://doi.org/doi:10.1016/j.future.2021.04.015

ASTEE. (2015). Gestion patrimoniale des réseaux d'assainissement. https://www.astee.org/

Beeri, C., Kanza, Y., Safra, E., & Sagiv, Y. (2004). Object fusion in geographic information systems. *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 816–827.

Belda, S., Pipia, L., Morcillo-Pallarés, P., Rivera-Caicedo, J. P., Amin, E., De Grave, C., & Verrelst, J. (2020). DATimeS: A machine learning time series GUI toolbox for gap-filling and vegetation phenology trends detection. *Environmental Modelling & Software*, *127*, 104666. https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104666

Belghaddar, Y., Seriai, A., Begdouri, A., Delenne, C., Chahinian, N., Bachar, R., & Derras, M. (2022). Towards a generic fusion framework for underground networks involving model-driven engineering domain [<hal-03685139>]. *International Journal of Information Science & Technology*. https://hal.archives-ouvertes.fr/hal-03685139

Belghaddar, Y., Chahinian, N., Seriai, A., Begdouri, A., Abdou, R., & Delenne, C. (2021). Graph convolutional networks: Application to database completion of wastewater networks. *Water*, *13*(12), 1681.

Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C.-H. (2014). Handling missing data in rcts; a review of the top medical journals. *BMC medical research methodology*, *14*(1), 1–8.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Bernold, L., Venkatesan, L., & Suvarna, S. (2003). A multi-sensory approach to 3-d mapping of underground utilities. *NIST SPECIAL PUBLICATION SP*, 525–530.

Bertolino, A., Calabrò, A., Lonetti, F., Di Marco, A., & Sabetta, A. (2011). Towards a model-driven infrastructure for runtime monitoring. In E. A. Troubitsyna (Ed.), *Software engineering for resilient systems* (pp. 130–144). Springer Berlin Heidelberg.

Bilal, M., Khan, W., Muggleton, J., Rustighi, E., Jenks, H., Pennock, S. R., Atkins, P. R., & Cohn, A. (2018). Inferring the most probable maps of underground utilities using bayesian mapping model. *Journal of Applied Geophysics*, *150*, 52–66. https://doi.org/https://doi.org/10.1016/j.jappgeo.2018.01.006

Bilenko, M., & Mooney, R. J. (2003). Employing trainable string similarity metrics for information integration. *IIWeb*, 67–72.

Bipartisan Infrastructure Law: State Revolving Funds Implementation Memorandum. (2022).

Bischof, S., Harth, A., Kämpgen, B., Polleres, A., & Schneider, P. (2018). Enriching integrated statistical open city data by combining equational knowledge and

missing value imputation. *Journal of Web Semantics*, *48*, 22–47. https://doi.org/10.1016/j.websem.2017.09.003

Boller, D., Moy de Vitry, M., D. Wegner, J., & Leitão, J. P. (2019). Automated localization of urban drainage infrastructure from public-access street-level images. *Urban Water Journal*, *16*(7), 480–493.

Bordogna, G., Pagani, M., & Pasi, G. (2010). Imperfect multisource spatial data fusion based on a local consensual dynamics. *Uncertainty approaches for spatial data modeling and processing* (pp. 79–94). Springer. https://doi.org/doi:10.1007/978-3-642-10663-7_6

Boury-Brisset, A.-C. (2013). Managing semantic big data for intelligence. *Semantic Technologies for Intelligence, Defense, and Security*, 41–47.

Broere, W. (2016). Urban underground space: Solving the problems of today's cities. *Tunnelling and Underground Space Technology*, *55*, 245–248.

Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2014). Spectral networks and locally connected networks on graphs.

Cai, H., Zheng, V. W., & Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), 1616–1637.

Canadian Common Ground Alliance. (2019). DIRT Report 2019.

Castanedo, F. (2013). A review of data fusion techniques. *The scientific world journal*, *2013*.

Chahinian, N., Bonnabaud La Bruyère, T., Frontini, F., Delenne, C., Julien, M., Panckhurst, R., Roche, M., Sautot, L., Deruelle, L., & Teisseire, M. (2021). Weir-p: An information extraction pipeline for the wastewater domain. *International Conference on Research Challenges in Information Science*, 171–188.

Chahinian, N., Delenne, C., Commandre, B., Derras, M., Deruelle, L., & Bailly, J.-S. (2019). Automatic mapping of urban wastewater networks based on manhole cover locations. *Computers, Environment and Urban Systems*, *78*, 101370.

Chen, H., & Cohn, A. G. (2011). Buried utility pipeline mapping based on multiple spatial data sources: A bayesian data fusion approach. *Twenty-Second International Joint Conference on Artificial Intelligence*.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th International Conference on Machine Learning*, 160–167. https://doi.org/10.1145/1390156.1390177

Commandre, B., En-Nejjary, D., Pibre, L., Chaumont, M., Delenne, C., & Chahinian, N. (2017). Manhole Cover Localization in Aerial Images with a Deep Learning Approach. *ISPRS - International Archives of the Photogrammetry, Remote*

Sensing and Spatial Information Sciences, *42W1*, 333–338. https://doi.org/ https://doi.org/10.5194/isprs-archives-XLII-1-W1-333-2017

Congressional Research Service. (2022). DOT's Federal Pipeline Safety Program: Background and Key Issues for Congress.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Costes, B. (2014). Matching Old Hydrographic Vector Data from Cassini's Maps. *e-Perimetron*, *9*(2), 51–65.

Costes, B., & Perret, J. (2019). A hidden markov model for matching spatial networks. *Journal of Spatial Information Science*, *2019*(18), 57–89. https://doi.org/doi: 10.5311/JOSIS.2019.18.489

COVADIS. (2019). *Standard de données réseaux d'AEP & d'assainissement, version 1.2* [Accessed 01 August 2020].

Da Silva, A. R. (2015). Model-driven engineering: A survey supported by the unified conceptual model. *Computer Languages, Systems & Structures*, *43*, 139–155.

Dasarathy, B. V. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, *85*(1), 24–38.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2017). Convolutional neural networks on graphs with fast localized spectral filtering.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. Retrieved August 10, 2022, from http: //www.jstor.org/stable/2984875

Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 325–339.

Deng, Y., Luo, A., Liu, J., & Wang, Y. (2019). Point of interest matching between different geospatial datasets. *ISPRS International Journal of Geo-Information*, *8*(10), 435. https://doi.org/doi:10.3390/ijgi8100435

Department for Environment, Food & Rural Affairs. (2002). Sewage treatment in the uk.

Devogele, T., Parent, C., & Spaccapietra, S. (1998). On spatial database integration. *International Journal of Geographical Information Science*, *12*(4), 335–352. https://doi.org/doi:10.1080/136588198241824

Dong, X. L., & Srivastava, D. (2013). Big data integration. *2013 IEEE 29th international conference on data engineering (ICDE)*, 1245–1248.

Du, J., Zhang, S., Wu, G., Moura, J. M., & Kar, S. (2017). Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*.

Dubois, D., & Prade, H. (2009). Formal representations of uncertainty. *Bouyssou d, dubois d, pirlot m, prade h, eds. decision-making process-concepts and methods* (Chapter 3. 85–156). ISTE & Wileyr.

Ducasse, S., Gîrba, T., Lanza, M., & Demeyer, S. (2005). Moose: A collaborative and extensible reengineering environment. *Tools for Software Maintenance and Reengineering*.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints.

El Faouzi, N.-E., Leung, H., & Kurian, A. (2011). Data fusion in intelligent transportation systems: Progress and challenges–a survey. *Information Fusion*, *12*(1), 4–10.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 1–37.

Environmental Protection Agency, Storm Water Management Model. (2022). *Epaswmm* [Accessed 1 august 2022].

Erraissi, A., & Belangour, A. (2018). Data sources and ingestion big data layers: Meta-modeling of key concepts and features. *International Journal of Engineering & Technology*, *7*(4), 3607–3612.

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., & Yin, D. (2019). Graph neural networks for social recommendation. *The world wide web conference*, 417–426.

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *37*(5), 692–709.

Fey, M., & Lenssen, J. E. (2019). Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*.

Gan, X., Liew, A. W.-C., & Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Research*, *34*(5), 1608–1619.

García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications*, *19*(2), 263–282. https://doi.org/10.1007/s00521-009-0295-6

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation [Advances in Machine Learning and Computational Intelligence]. *Neurocomputing*, *72*(7), 1483–1493. https://doi.org/https://doi.org/10.1016/j.neucom.2008.11.026

Gascueña, J. M., Navarro, E., & Fernández-Caballero, A. (2012). Model-driven engineering techniques for the development of multi-agent systems. *Engineering Applications of Artificial Intelligence*, *25*(1), 159–173.

Geoff, Z., & Sakura, S. (2020). Reducing Damage to Underground Utility Infrastructure during Excavation.

Giustarini, L., Parisot, O., Ghoniem, M., Hostache, R., Trebs, I., & Otjacques, B. (2016). A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records. *Environmental Modelling & Software*, *82*, 308–320. https://doi.org/10.1016/j.envsoft.2016.04.013

Goodwin, P. (2005). Utilities' street works and the cost of traffic congestion.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864.

Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 922–929.

Hafsi, M., Bolon, P., & Dapoigny, R. (2017). Detection and localization of underground networks by fusion of electromagnetic signal and GPR images. In H. Nagahara, K. Umeda, & A. Yamashita (Eds.), *Thirteenth international conference on quality control by artificial vision 2017* (pp. 7–14). SPIE. https://doi.org/10.1117/12.2266946

Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, *85*(1), 6–23.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 1024–1034.

Hammond, D. K., Vandergheynst, P., & Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, *30*(2), 129–150. https://doi.org/https://doi.org/10.1016/j.acha.2010.04.005

Hamzah, F. B., Mohd Hamzah, F., Mohd Razali, S. F., Jaafar, O., & Abdul Jamil, N. (2020). Imputation methods for recovering streamflow observation: A methodological review. *Cogent Environmental Science*, *6*(1), 1745133.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, J., Keung, J. W., Sarro, F., Li, Y.-F., Yu, Y.-T., Chan, W., & Sun, H. (2017). Cross-validation based k nearest neighbor imputation for software quality datasets: An empirical study. *Journal of Systems and Software*, *132*, 226–252.

Ineris [Accessed 20 september 2022]. (2022).

INSEE. (2019). https://www.insee.fr

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913–933.

Jepsen, T. S., Jensen, C. S., & Nielsen, T. D. (2019). Graph convolutional networks for road networks. *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. https://doi.org/10.1145/3347146.3359094

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, *50*(2), 105–115.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2020). Handling incomplete and missing data in water network database using imputation methods. *Sustainable and Resilient Infrastructure*, *5*(6), 365–377.

Kamkhad, N., Jampachaisri, K., Siriyasatien, P., & Kesorn, K. (2020). Toward semantic data imputation for a dengue dataset. *Knowledge-Based Systems*, *196*, 105803.

Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, *14*(1), 28–44.

Kim, J. O., Yu, K., Heo, J., & Lee, W. H. (2010). A new method for matching objects in two different geospatial datasets based on the geographic context. *Computers & Geosciences*, *36*(9), 1115–1122. https://doi.org/doi:10.1016/j.cageo.2010.04.003

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks.

Kofinas, D. T., Spyropoulou, A., & Laspidou, C. S. (2018). A methodology for synthetic household water consumption data generation. *Environmental Modelling & Software*, *100*, 48–66. https://doi.org/https://doi.org/10.1016/j.envsoft.2017.11.021

Koschmann, M., Collins, L., Spencer, T., & Moon, S. (2021). Empirical investigation into underground utility strikes for supporting incident prevention: Cases in melbourne, australia. *International Journal of Occupational Safety and Ergonomics*, 1–9.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kumar, A., Rizvi, S. M. A. A., Brooks, B., Vanderveld, R. A., Wilson, K. H., Kenney, C., Edelstein, S., Finch, A., Maxwell, A., Zuckerbraun, J., & Ghani, R. (2018). Using machine learning to assess the risk of and prevent water main breaks. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

la Fosse, T. B., Cheng, Z., Rocheteau, J., & Mottu, J. .-. (2020). Model-driven engineering of monitoring application for sensors and actuators networks. *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 553–560. https://doi.org/https://doi.org/10.1109/SEAA51224.2020.00091

Legifrance [Accessed 01 august 2020]. (2012).

Li, L., & Goodchild, M. F. (2011). An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, *2*(4), 309–328. https://doi.org/doi:10.1080/19479832.2011.577458

Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2017). Gated graph sequence neural networks.

Liew, A. W.-C., Law, N.-F., & Yan, H. (2010). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, *12*(5), 498–513. https://doi.org/10.1093/bib/bbq080

Lin, P., & Yuan, X. X. (2019). A two-time-scale point process model of water main breaks for infrastructure asset management. *Water Research*, *150*, 296–309. https://doi.org/10.1016/j.watres.2018.11.066

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

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Little, T., Jorgensen, T., Lang, K., & Moore, E. W. (2013). On the joys of missing data. *Journal of pediatric psychology*, *39*. https://doi.org/10.1093/jpepsy/jst048

Ma, J., Cheng, J. C., Ding, Y., Lin, C., Jiang, F., Wang, M., & Zhai, C. (2020). Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series. *Advanced Engineering Informatics*, *44*(March), 101092. https://doi.org/10.1016/j.aei.2020.101092

Mapping the Underworld [Accessed: 2022-05-12]. (2022).

McMahon, W., Burtwell, M., & Evans, M. (2005). Minimising street works disruption: The real costs of street works to the utility industry and society. london: Uk water industry research.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Min, D., Zhilin, L., & Xiaoyong, C. (2007). Extended hausdorff distance for spatial objects in gis. *International Journal of Geographical Information Science*, *21*(4), 459–475. https://doi.org/doi:10.1080/13658810601073315

Montpellier Méditeranée Métropole. (2020). *Open data* [Accessed 1 august 2020].

Moose. (2022).

Muggleton, J., & Rustighi, E. (2013). 'mapping the underworld': Recent developments in vibro-acoustic techniques to locate buried infrastructure. *Géotechnique Letters*, *3*(3), 137–141.

Murtojärvi, M., Suominen, T., Uusipaikka, E., & Nevalainen, O. S. (2011). Optimising an observational water monitoring network for Archipelago Sea, South West Finland. *Computers and Geosciences*, *37*(7), 844–854. https://doi.org/10.1016/j.cageo.2011.01.006

Nassreddine, G., Abdallah, F., & Denoeux, T. (2009). Map matching algorithm using interval analysis and dempster-shafer theory. *2009 IEEE Intelligent Vehicles Symposium*, 494–499. https://doi.org/doi:10.1109/IVS.2009.5164328

Nelwamondo, F. V., Golding, D., & Marwala, T. (2013). A dynamic programming approach to missing data estimation using neural networks. *Information Sciences*, *237*, 49–58. https://doi.org/10.1016/j.ins.2009.10.008

Nelwamondo, F. V., Mohamed, S., & Marwala, T. (2007). Missing data: A comparison of neural network and expectation maximization techniques. *Current Science*, 1514–1521.

Ngouna, R. H., Ratolojanahary, R., Medjaher, K., Dauriac, F., Sebilo, M., & Junca-Bourié, J. (2020). A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values. *Engineering Applications of Artificial Intelligence*, *95*(July), 103822. https://doi.org/10.1016/j.engappai.2020.103822

Nishanth, K. J., & Ravi, V. (2016). Probabilistic neural network based categorical data imputation. *Neurocomputing*, *218*, 17–25.

Object Management Group. (2006). *Meta object facility (MOF) 2.0 core specification* [Accessed 1 august 2020].

Open platform for french public data [Accessed 01 august 2020]. (2015).

Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access*, *6*, 63279–63291.

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., & Belfkih, S. (2018). Big data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, *30*(4), 431–448.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. *NIPS 2017 Workshop on Autodiff*. https://openreview.net/forum?id=BJJsrmfCZ

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, *9*, 157.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pharo. (2022).

Rahimi, A., Cohn, T., & Baldwin, T. (2018). Semi-supervised user geolocation via graph convolutional networks.

Raimond, A.-M. O., Mustière, S., & Ruas, A. (2015). Knowledge formalization for vector data matching using belief theory. *J. Spatial Inf. Sci.*, *10*, 21–46. https://doi.org/doi:10.5311/JOSIS.2015.10.194

Raol, J. R. (2015). *Data fusion mathematics: Theory and practice*. CRC Press.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *ArXiv*, *abs/1804.02767*.

Rosen, B., & Saalfeld, A. (1985). Match criteria for automatic alignment. *Proceedings of 7th international symposium on computer-assisted cartography (Auto-Carto 7)*, 1–20.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rucklidge, W. (1996). *Efficient visual recognition using the hausdorff distance*. Springer-Verlag.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Saalfeld, A. (1988). Conflation automated map compilation. *International Journal of Geographical Information System*, *2*(3), 217–228. https://doi.org/doi:10.1080/02693798808927897

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(3), 660–674. https://doi.org/10.1109/21.97458

Samal, A., Seth, S., & Cueto 1, K. (2004). A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, *18*(5), 459–489. https://doi.org/doi:10.1080/13658810410001658076

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of Mean Values and covariance matrices and imputation of Missing values. *Journal of Climate*, *14*(5), 853–871. https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

Serrano-Notivoli, R., de Luis, M., & Beguería, S. (2017). An R package for daily precipitation climate series reconstruction. *Environmental Modelling & Software*, *89*, 190–195. https://doi.org/https://doi.org/10.1016/j.envsoft.2016.11.005

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton university press.

Simone, G., Farina, A., Morabito, F. C., Serpico, S. B., & Bruzzone, L. (2002). Image fusion techniques for remote sensing applications. *Information fusion*, *3*(1), 3–15.

Smets, P. (1997). Imperfect information: Imprecision and uncertainty. *Uncertainty management in information systems* (pp. 225–254). Springer.

Smets, P., & Kennes, R. (1994). The transferable belief model. *Artificial intelligence*, *66*(2), 191–234. https://doi.org/doi:10.1016/0004-3702(94)90026-4

Smith, B. L., Scherer, W. T., & Conklin, J. H. (2003). Exploring imputation techniques for missing data in transportation management systems. *Transportation Research Record*, *1836*(1), 132–142.

Smith, D., & Singh, S. (2006). Approaches to multisensor data fusion in target tracking: A survey. *IEEE transactions on knowledge and data engineering*, *18*(12), 1696–1710.

Song, W., Keller, J. M., Haithcoat, T. L., & Davis, C. H. (2011). Relaxation-based point feature matching for vector map conflation. *Transactions in GIS*, *15*(1), 43–60. https://doi.org/doi:10.1111/j.1467-9671.2010.01243.x

Spinelli, I., Scardapane, S., & Uncini, A. (2020). Missing data imputation with adversarially-trained graph convolutional networks. *Neural Networks*, *129*, 249–260. https://doi.org/10.1016/j.neunet.2020.06.005

Strahler, A. (1957). Quantitative analysis of watershed geomorphology. *Eos, Transactions American Geophysical Union*, *38*(6), 913–920. https://doi.org/https://doi.org/10.1029/TR038i006p00913

Sun, Z., Wang, P., Vuran, M. C., Al-Rodhaan, M. A., Al-Dhelaan, A. M., & Akyildiz, I. F. (2011). Mise-pipe: Magnetic induction-based wireless sensor networks for underground pipeline monitoring. *Ad Hoc Networks*, *9*(3), 218–227.

Thekumparampil, K. K., Wang, C., Oh, S., & Li, L.-J. (2018). Attention-based graph neural network for semi-supervised learning.

Tong, X., Liang, D., & Jin, Y. (2014). A linear road object matching method for conflation based on optimization and logistic regression. *International Journal of Geographical Information Science*, *28*(4), 824–846. https://doi.org/doi:10.1080/13658816.2013.876501

Tong, X., Shi, W., & Deng, S. (2009). A probability-based multi-measure feature matching method in map conflation. *International Journal of Remote Sensing*, *30*(20), 5453–5472. https://doi.org/doi:10.1080/01431160903130986

Tsai, C.-F., & Chang, F.-Y. (2016). Combining instance selection for better missing value imputation. *Journal of Systems and Software*, *122*, 63–71. https://doi.org/https://doi.org/10.1016/j.jss.2016.08.093

Tsiami, L., & Makropoulos, C. (2021). Cyber—physical attack detection in water distribution systems with temporal graph convolutional neural networks. *Water*, *13*(9). https://doi.org/10.3390/w13091247

United Nations and Department of Economic and Social Affairs and Population Division. (2019). World urbanization prospects: The 2018 revision.

USAG Data and Reporting Working Group. (2019). Utility Strike Damages Report.

Volz, S. (2006). 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.

Wald, L. (1999). Some terms of reference in data fusion. *IEEE Transactions on geoscience and remote sensing*, *37*(3), 1190–1193.

Walter, V., & Fritsch, D. (1999). Matching spatial data sets: A statistical approach. *International Journal of geographical information science*, *13*(5), 445–473. https://doi.org/doi:10.1080/136588199241157

Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J., & Zhang, Z. (2019). Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

Wang, S., Guo, Q., Xu, X., & Xie, Y. (2019). A study on a matching algorithm for urban underground pipelines. *ISPRS International Journal of Geo-Information*, *8*(8), 352. https://doi.org/doi:10.3390/ijgi8080352

Wang, Y., Chen, D., Zhao, Z., Ren, F., & Du, Q. (2015). A back-propagation neural network-based approach for multi-represented feature matching in update propagation. *Transactions in GIS*, *19*(6), 964–993. https://doi.org/doi:10.1111/tgis.12138

White, F. (1987). Joint directors of laboratories-technical panel for c3i, data fusion sub-panel. *San Diego: Naval Ocean Systems Center*.

Wu, J., Wan, Y., Chiang, Y.-Y., Fu, Z., & Deng, M. (2018). A matching algorithm based on voronoi diagram for multi-scale polygonal residential areas. *IEEE Access*, *6*, 4904–4915. https://doi.org/doi:10.1109/ACCESS.2018.2793302

Xavier, E. M., Ariza-López, F. J., & Urena-Camara, M. A. (2016). A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys (CSUR)*, *49*(2), 1–34. https://doi.org/doi:10.1145/2963147

Xu, D., Hu, P. J.-H., Huang, T.-S., Fang, X., & Hsu, C.-C. (2020). A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management. *Journal of Biomedical Informatics*, *111*, 103576.

Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, *160*(April), 104–118. https://doi.org/10.1016/j.knosys.2018.06.012

Young, W., Weckman, G., & Holland, W. (2011). A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits. *Theoretical Issues in Ergonomics Science*, *12*(1), 15–43.

Zhang, M., Shi, W., & Meng, L. (2005). A generic matching algorithm for line networks of different resolutions. *Workshop of ICA commission on generalization and multiple representation computering faculty of a Coruña University-Campus de Elviña, Spain*, *9*, 101–110.

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2019). Graph neural networks: A review of methods and applications.