كلية العلوم و التقنيات فاس
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
Université Sidi Mohamed Ben Abdellah

Département d'Informatique

Projet de fin d'études

Master Sciences et Techniques
Systèmes Intelligents & Réseaux

Fusion de données spatiales : Gestion des données imparfaites en utilisant les fonctions de croyance

HydroSciences Montpellier

Lieu du stage : Laboratoire Hydrosciences / Equipe Inria Lemon

Réalisé par :

- ET-TARGUY Omar

Encadré par :

- Pr. Ahlame BEGDOURI
- Dr. Carole DELENNE
- Pr. Nanée CHAHINIAN
- Mr. Yassine BELGHADDAR

Soutenu le 15.07.2021 devant le jury composé de :

- Pr. Azeddine ZAHI	Faculté des Sciences et Techniques de Fès	(Président)
- Pr. Ilham CHAKER	Faculté des Sciences et Techniques de Fès	(Examinateur)
- Pr. Ahlame BEGDOURI	Faculté des Sciences et Techniques de Fès	(Encadrante)
- Dr. Carole DELENNE	HydroSciences Montpellier	(Encadrante)
- Mr. Yassine BELGHADDAR	HydroSciences Montpellier/ Faculté des Sciences et Techniques de Fès /Berger-Levrault	(Encadrant)

Année Universitaire 2020– 2021

Remerciements

Avant toute chose, je tiens à remercier tous ceux qui ont participé de près ou de loin à la réussite de ce stage et à la réalisation ce rapport.

Je tiens à remercier en premier lieu **Pr. Ahlame BEGDOURI**, **Pr. Carole DELENNE**, **Pr. Nanné CHAHINIAN** et **Mr. Yassine BEL-GHADDAR**, mes encadrants durant ce stage de fin d'études, pour l'orientation, l'aide et les conseils qui m'ont guidé durant l'élaboration de ce travail, ainsi que pour le temps consacré aux différentes réunions qui ont rythmé toute la période de ce stage.

Mes remerciements vont également aux membres du jury, **Pr. Ilham CHAKER** et **Pr. Azeddine ZAHI**, pour l'intérêt qu'ils ont porté à ce mémoire et pour avoir accepté de l'évaluer et de l'enrichir par leurs propositions.

Je remercie le corps professoral de la filière Systèmes Intelligents et Réseaux pour la formation reçue. Merci à chacun d'eux pour le transfert de savoir.

Enfin, un merci particulier à mes parents et à mes amis pour m'avoir toujours soutenu lors de mes études.

Résumé

De nos jours, les données concernant les réseaux d'assainissement qui couvrent le même territoire sont disponibles à partir de plusieurs sources. Par exemple, les nœuds des réseaux d'assainissement (regards, avaloirs, etc.) sont accessibles à partir de la source officielle de données géographiques de Montpellier 3M[1] et d'images à haute résolution. Cette diversité de sources a donné lieu à plusieurs travaux axés sur le thème de la visualisation, l'analyse et la fusion de bases de données spatiales à des fins diverses telles que l'amélioration de la qualité ou la fourniture de nouveaux services.

Pour rendre plus explicites les relations entre les différents objets des bases de données, nous utilisons le processus d'appariement des données géographiques qui consiste à mettre en correspondance les objets homologues représentant la même réalité.

La complexité du processus d'appariement implique que les approches existantes varient en fonction des besoins que la correspondance doit satisfaire, et varient en fonction des types de données à mettre en correspondance (points, lignes ou surfaces) et du degré de détail. Nous avons trouvé que la plupart des approches sont basées sur la géométrie et leurs relations et que très peu prennent en charge les connaissances sémantiques.

Pour tenir compte de tous les critères géométriques et sémantiques et leurs relations, nous utilisons la théorie des fonctions de croyance, dans le but de formaliser la connaissance et de modéliser les imperfections des données rencontrées lors de la fusion des données des réseaux d'assainissement multi-source.

Mots-clés : théorie des fonctions de croyance, appariement, fusion, connaissances, imperfection.

1. *Montpellier Méditerranée Métropole*

Abstract

Nowadays, data concerning sewerage networks covering the same territory are available from various sources. For example, the nodes of sewerage networks (manholes, gullies, etc.) are accessible from the official Montpellier geographical data source 3M and from high-resolution images. This diversity of sources has given rise to several works focusing on the visualisation, analysis and fusion of spatial databases for various purposes such as quality enhancement or the provision of new services.

To make the relationships between the different objects in the databases more explicit, we use the process of geographic data matching, which consists in matching homologous objects representing the same reality.

The complexity of the matching process means that existing approaches vary according to the needs that the matching must satisfy, and vary according to the kinds of data to be matched (points, lines or surfaces) and the degree of detail. We have found that most approaches are based on geometry and its relationships and very few supports semantic knowledge.

In order to take into account all the geometric and semantic criteria and their relations, we use the belief functions theory, with the aim of formalising the knowledge and modelling the data imperfections encountered when fusing data from multi-source sewerage networks.

Keywords : belief functions theory, matching, fusion, knowledge, imperfection.

Table des matières

Remerciements **i**

Résumé **ii**

Abstract **iii**

Introduction générale **1**

1 Cadre général du projet **3**

1.1 Établissement d'accueil 4

1.2 Problématique 5

1.3 Contribution 6

2 Théorie des fonctions de croyance **7**

2.1 Introduction 8

2.2 Niveau crédal statique : Représentation des informations 9

2.2.1 Cadre de discernement 9

2.2.2 Fonction de masse 9

2.2.3 Fonctions de masse particulaires 10

2.2.4 Autres fonctions de croyance 11

2.2.5 Opérations sur la représentation des connaissances : raffinement / grossissement 12

2.3 Niveau crédal dynamique : manipulation des informations 13

2.3.1 Conditionnement et déconditionnement 13

2.3.2 Affaiblissement 14

2.3.3 Combinaison de fonction de croyance 14

2.4 Niveau pignistique : Prise de décision 16

2.5 Gestion du confit 16

2.6 Conclusion 17

3 Appariement des objets géographiques **18**

3.1 Définition 19

3.2 Les mesures de similarité 20

3.2.1 Mesures Géométrique 21

3.2.1.1 Position 21

3.2.1.2 Forme 22

3.2.2 Mesures Topologiques 22

3.2.2.1 Degré d'un noeud 22

3.2.2.2 Direction des lignes 23

3.2.2.3 Autres 23

3.2.3 Mesures Attributaires 25

3.3 Les méthodes d'appariements 27

3.3.1 Classification selon le nombre de mesures de similarités 27

3.3.2 Classification selon la relation des méthodes avec les mesures 27

3.3.3 Classification selon la direction d'appariement 28

3.3.4 Classification selon les étapes d'appariement 28

3.3.5 Classification Statistique / non-statistique 32

3.3.6 Classification Itérative / non-itérative 32

4 Contribution à la fusion des données imparfaites **33**

4.1 Introduction 34

4.2 Catégorisations des formes d'imperfections liées aux réseaux d'assainissement . 34

4.2.1 Formes d'imperfections liées aux nœuds 34

4.2.1.1 Décalage d'un nœud avec un seul candidat 34

4.2.1.2 Décalage d'un nœud avec plusieurs candidats 35

4.2.2 Formes d'imperfections liées aux arêtes 36

4.2.2.1 Décalage uniforme de type 1 :1 36

4.2.2.2 Décalage non uniforme de type 1 :1 37

4.2.2.3 Décalage uniforme de type 1 :n 37

4.2.2.4 Décalage non uniforme de type 1 :n 38

4.3 Le processus d'appariement de données géographiques selon [29] 38

4.3.1 Choix de l'approche 38

4.3.2 Principe général . . . 39
4.3.3 Sélection des candidats . . . 40
4.3.4 Initialisation des masses de croyance . . . 41
4.3.5 Fusion des critères . . . 42
4.3.6 Fusion des candidats . . . 43
4.3.7 Décision . . . 44
4.4 Adaptation de l'approche à notre contexte . . . 45
4.4.1 Les bases de données utilisées . . . 45
4.4.2 Initialisation des masses de croyance . . . 46
4.4.2.1 La géométrie des objets géographiques . . . 47
4.4.2.2 Le critère de la distance euclidienne . . . 47
4.4.2.3 Le critère de l'angle . . . 48
4.4.2.4 Le critère de la distance de Hausdorff . . . 50
4.4.2.5 Le critère du score . . . 51
4.4.2.6 Le critère de la longueur . . . 52
4.5 Non fiabilité des sources . . . 53

5 Expérimentations et résultats **55**
5.1 Scénario de l'expérimentation . . . 56
5.2 Expérimentations pour les nœuds . . . 56
5.2.1 Appariement d'un nœud : cas d'un seul candidat potentiel . . . 57
5.2.2 Appariement d'un nœud : cas de plusieurs candidats potentiels . . . 59
5.3 Expérimentations pour les arêtes . . . 61
5.3.1 Décalage uniforme de type 1 : 1 . . . 62
5.3.2 Décalage non uniforme de type 1 : 1 . . . 64
5.3.3 Décalage uniforme de type 1 : n . . . 67
5.3.4 Décalage non uniforme de type 1 : n . . . 72
5.4 Synthèse et discussion . . . 74

Conclusions et perspectives **76**

Annexes **84**

A Définitions **85**

A.1 Les formes d'imperfection d'information 85
A.2 QGIS 86
A.3 Les types de données SIG 86
A.3.1 Les données Raster 86
A.3.2 Les données Vecteur 87
A.4 Le stockage de données SIG 88
A.5 GDAL 89

B Les formalismes de gestion des données imparfaites 90

B.1 La théorie des probabilités 90
B.2 La théorie des ensembles approximatifs 91
B.3 La théorie des ensembles flous 91
B.4 La théorie des possibilités 92
B.4.1 La mesure de possibilité 92
B.4.2 La mesure de nécessité 93
B.5 La théorie des fonctions de croyance 94

Table des figures

1.1 Organigramme HSM 4

2.1 Schéma des deux niveaux du Modèle des Croyances Transférables MCT[2] [1] . . 8

2.2 Fonctions de masses particulaires 10

2.3 Exemple de crédibilité 11

2.4 Exemple de plausibilité 11

2.5 Grossissement / Raffinement 12

3.1 Visualisation de deux source de données spatiales 19

3.2 Ingrédients d'appariement d'objets. 20

3.3 La distance Hausedouff entre deux lignes L1 et L2 [57] 21

3.4 Calcul de la forme 22

3.5 Distribution de probabilité de la forme 22

3.6 Probabilités conditionnelles de la forme 22

3.7 Degré d'un noeud [63] 23

3.8 Discrétisation de l'orientation des lignes selon 8 axes [32] 23

3.9 relations topologiques des voisins [4] 24

3.10 Graphe de proximité [20] 24

3.11 Introduction d'un nouveau critère basé sur la distance aux entités ponctuelles déjà appariées [9] 25

3.12 Buffet de sélection 29

3.13 Matrice de similarités [36] 30

3.14 Buffer de sélection 30

3.15 Graphe de similarité entre trois sources [36] 31

4.1 Exemple de décalage d'un nœud avec un seul candidat 35

2. *Modèle des Croyances Transférables*

4.2 Exemple de décalage d'un nœud avec plusieurs candidats 35
4.3 Exemple de décalage uniforme entre les lignes avec une cardinalité de 1 :1 36
4.4 Exemple de décalage non uniforme entre les lignes avec une cardinalité de 1 :1 . 37
4.5 Exemple de décalage uniforme entre les lignes avec une cardinalité de 1 :n 37
4.6 Exemple de décalage non uniforme entre les lignes avec une cardinalité de 1 :n . 38
4.7 Processus d'appariement de données géographiques détaillé [29] 39
4.8 Etape de sélection des candidats à l'appariement 40
4.9 Candidats à l'appariement pour l'objet 525 sélectionnés selon un critère de distance géométrique 40
4.10 Deuxième étape du processus : initialisation des masses de croyance 42
4.11 Troisième étape du processus : fusion des critères 43
4.12 Fusion des critères pour le candidat C_i 43
4.13 Quatrième étape du processus : fusion des candidats 43
4.14 Combinaison des masses de croyance pour chacun des candidats 44
4.15 Cinquième étape du processus : décision 44
4.16 Décision selon le maximum de probabilité pignistique 45
4.17 Exemple de noeuds détectés à partir d'images à haute résolution 46
4.18 Représentation des connaissances pour le critère de la distance euclidienne . . . 47
4.19 Produit scalaire entre deux vecteurs 49
4.20 Représentation des connaissances pour le critère de l'angle 49
4.21 Représentation des connaissances pour le critère de la distance de Hausdorff . . 50
4.22 Représentation des connaissances pour le critère du score 51
4.23 Représentation des connaissances pour le critère de la longueur 52
4.24 Exemple d'affaiblissement 53
4.25 Exemple des masses attribuées aux candidats de l'arête numéro 99 en fonction de la distance de hausdorff 54

5.1 Décalage d'un nœud avec un seul candidat 57
5.2 Sélection des candidats du nœud numéro 648 57
5.3 Initialisation des masses de croyance pour chaque critère 57
5.4 Jeu de masses après la combinaison des mesures 58
5.5 Jeux de masse après la normalisation 58
5.6 Probabilité pignistique pour chaque hypothèse 59

5.7 Décalage d'un nœud avec plusieurs candidats 59
5.8 Sélection des candidats du nœud numéro 495 59
5.9 Initialisation des masses de croyance pour chaque critère 60
5.10 Jeu de masses après la combinaison des mesures 60
5.11 Fusion des candidats 61
5.12 Probabilité pignistique pour chaque hypothèse du cadre de discernement 61
5.13 Exemple de décalage uniforme avec une cardinalité de 1 : 1 62
5.14 Sélection des candidats de l'arête numéro 898 62
5.15 Initialisation des masses de croyance pour chaque critère 63
5.16 Jeu de masses après la combinaison des mesures 63
5.17 Fusion des candidats 64
5.18 Probabilité pignistique pour chaque hypothèse du cadre de discernement 64
5.19 Exemple de décalage non uniforme avec une cardinalité de 1 : 1 65
5.20 Sélection des candidats de l'arête numéro 99 65
5.21 Initialisation des masses de croyance pour chaque critère 65
5.22 Jeu de masses après la combinaison des mesures 66
5.23 Fusion des candidats 66
5.24 Probabilité pignistique pour chaque hypothèse du cadre de discernement 67
5.25 Décalage uniforme avec une cardinalité de 1 : n 67
5.26 Sélection des candidats de l'arête numéro 897 68
5.27 Initialisation des masses de croyance pour chaque critère 68
5.28 Jeu de masses après la combinaison des mesures 68
5.29 Probabilité pignistique pour chaque hypothèse du cadre de discernement 69
5.30 Exemple des strokes 70
5.31 Décalage uniforme avec une cardinalité de 1 : n 70
5.32 Détection des strokes 71
5.33 Initialisation des masses de croyance pour chaque critère 71
5.34 Jeu de masses après la combinaison des mesures 72
5.35 Probabilité pignistique pour chaque hypothèse du cadre de discernement 72
5.36 Décalage non uniforme avec une cardinalité de 1 : n 72
5.37 Détection des strokes 73
5.38 Sélection des candidats du stroke numéro 1100 73

5.39 Initialisation des masses de croyance pour chaque critère 73
5.40 Jeu de masses après la combinaison des mesures 74
5.41 Probabilité pignistique pour chaque hypothèse du cadre de discernement 74
5.42 Variation du coefficient de normalisation en fonction du conflit [7] 75

A.1 Grands types d'imperfection 85
A.2 Exemple d'une couche Raster 87
A.3 Exemple de données Vecteur 87
A.4 Exemple d'une table attributaire 88

Liste d'acronymes et abréviations

HSM	*HydroSciences Montpellier*
CNRS	*Centre National de la Recherche Scientifique*
MCT	*Modèle des Croyances Transférables*
IRD	*Institut de Recherche pour le Développement*
UMR	*Unité Mixte de Recherche*
SIG	*Système d'Information Géographique*
QGIS	*Quantum Geographic Information System*
GDAL	*Geospatial Data Abstraction Library*
3M	*Montpellier Méditerranée Métropole*

Introduction générale

Depuis une dizaine d'années, le Big Data est l'une des principales innovations dans le domaine des systèmes d'information. Auparavant, la quantité de données disponibles était très limitée pour de nombreuses raisons. Par exemple, la taille insuffisante des dispositifs de stockage. De nos jours, et grâce au développement et à la croissance de l'utilisation des nouvelles technologies telles que l'Internet, les smartphones, les objets connectés et les GPS. Les données relatives à un domaine sont maintenant disponibles à travers plusieurs sources avec des masses énormes, ce qui a dépassé plusieurs obstacles liés à l'existence des données. D'autre part, cette explosion quantitative a donné lieu à de nouveaux problèmes liés à son traitement et à son exploitation.

Les données concernant les réseaux d'assainissement sont à priori disponibles à partir de plusieurs sources de données. Or, depuis leur construction, ces réseaux subissent fréquemment des travaux d'amélioration et d'extension. Les entreprises qui participent à leur maintenance sont multiples et changent régulièrement. Par conséquent, les cartes des réseaux disponibles émanant de sources différentes contiennent parfois des données identiques, complémentaires ou contradictoires.

La particularité des données issues de Systèmes d'Informations Géographiques (SIG) est qu'elles sont constituées à la fois, d'éléments géométriques, et d'une base de données d'attributs. La fusion des données spatiales multi-sources permet de créer un nouvel ensemble de données plus riche. Or, le plus souvent, ces informations sont imparfaites : imprécises, incertaines, vague, incomplètes, etc. Il est alors indispensable de recourir à des formalismes qui permettent de modéliser ces imperfections. Par ailleurs, afin de profiter au maximum de ces imperfections, ces formalismes doivent aussi proposer des mécanismes de fusion. Cette étape permet d'obtenir une information synthétique, qui résume l'ensemble des données initiales.

Le formalisme de la théorie des fonctions de croyance permet de représenter la connais-

sance d'une manière relativement naturelle et de modéliser les diverses formes d'imperfection de l'information. De plus, les bases de cette théorie sont très simples à comprendre, Cela représente un avantage important. De plus, cette théorie unifie toutes les théories de l'incertitude. Pour ces raisons, nous nous sommes essentiellement tournés vers ce formalisme.

Le reste de ce mémoire est organisé de la manière suivante :

Dans le chapitre 1, nous allons présenter le cadre général de ce projet, la problématique ainsi que les solutions proposées.

Dans le deuxième chapitre, nous présentons les différents formalismes de la gestion des données imparfaites. Ainsi, nous allons faire une revue de la littérature sur les différents travaux concernant la gestion des imperfections dans le cadre de la théorie des fonctions de croyance.

Le chapitre 3 est consacré aux principaux outils mathématiques proposés dans le cadre de la théorie des fonctions de croyance.

Dans le chapitre 4, nous présentons le processus et les méthodes d'appariement des objets géographiques.

Le chapitre 5 est consacré aux principaux outils informatiques et bibliothèques utilisés.

Dans le chapitre 6, nous catégorisons les formes d'imperfection liées aux réseaux d'assainissement. Ainsi, nous présentons l'approche d'appariement de données géographiques utilisée. Enfin, nous adaptons cette approche à notre contexte.

Le chapitre 7 est consacré aux expérimentations et résultats obtenus pour chaque catégorie d'imperfection.

Nous concluons ce rapport par une conclusion générale qui rappelle les grandes lignes de ce travail, la problématique, la démarche et les résultats ainsi qu'une discussion visant à présenter les perspectives et les améliorations futures.

Chapitre 1

Cadre général du projet

1.1 Établissement d'accueil

Le Laboratoire HydroSciences Montpellier [17] HSM [1] figure 1.1 est une Unité Mixte de Recherche de l'Université de Montpellier, de l'IRD [2] et du CNRS [3], qui dédie ses activités de recherche, d'enseignement et de formation à l'amélioration et à la transmission des connaissances en sciences de l'eau.

HydroSciences Montpellier comprend parmi ses personnels, près de 170 personnes, 120 permanents et une cinquantaine de doctorants et post-doctorants, localisés sur trois sites de l'Université de Montpellier (Maison des Sciences de l'Eau – campus Triolet et Faculté de Pharmacie) et au sein de certaines implantations IRD dans les pays du Sud. L'UMR [4] accueille également une centaine de stagiaires par an.

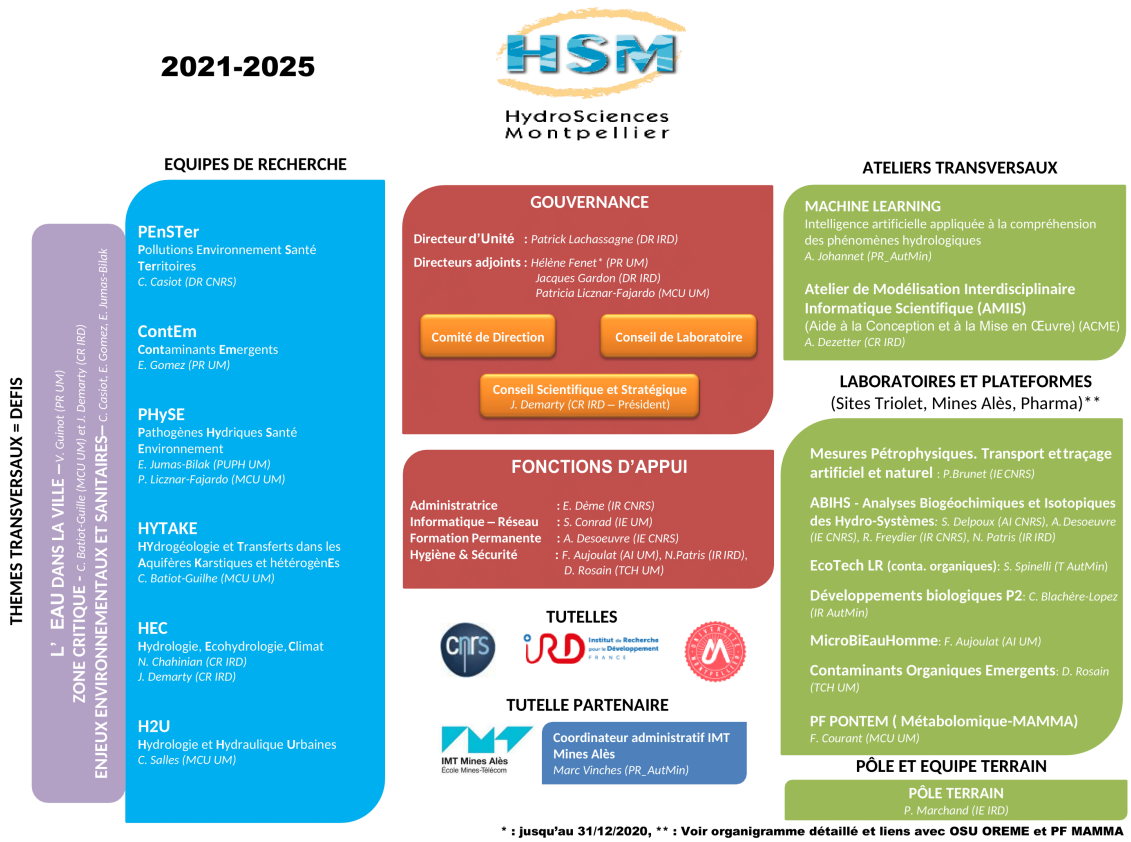


FIGURE 1.1 – Organigramme HSM

1. *HydroSciences Montpellier*
2. *Institut de Recherche pour le Développement*
3. *Centre National de la Recherche Scientifique*
4. *Unité Mixte de Recherche*

1.2 Problématique

Un réseau d'assainissement est un graphe constitué de nœuds et d'arêtes. Les nœuds représentent des ouvrages, des appareillages et des réparations (regards, avaloir etc.), les arêtes illustrent les canalisations (collecte, transport, etc.). Chacun des nœuds et des arêtes possède un ensemble de propriétés sous forme d'attributs. Nous divisons donc les données sur ces réseaux en deux catégories :

- Spatiale/Géométrique-Topologique : les points et les lignes (forme, position et leurs relations).
- Attributaire/sémantique : les attributs et leurs interprétations.

Cependant, ces données sont imparfaites : imprécises, incomplètes, incertaines, vagues, etc. l'imperfection de l'information fait appel à plusieurs concepts.

Le premier concept concerne l'imprécision des informations. Les imprécisions qui correspondants une difficulté dans l'énoncé de la connaissance, soit parce que des connaissances numériques sont mal connues, dans notre cas, par exemple, diamètre d'une conduite entre la valeur 150 et 200, soit parce que des termes du langage naturel sont utilisés pour qualifier certaines caractéristiques, par exemple, Les différents stades de la vie d'un Homme (Bébé, Enfant, Adolescent, adulte, personne âgée). Enfin la notion de l'imprécision fait appel à la notion d'incomplétude qui correspond à l'absence d'information, dans notre cas, par exemple, le champ type de matériel n'est pas renseigné pour une conduite.

Le deuxième concept concerne l'incertitude, qui est à différencier de l'imprécision, car il fait référence non pas au contenu de l'information mais à sa qualité. L'incertitude fait référence à la véracité de l'information (l'information peut être complète et précise mais fausse), dans notre cas, par exemple l'existence (ou non) d'un regard ou d'une conduite. Les données sur les mêmes objets réels sont accessibles à travers plusieurs sources. Notre problématique est la suivante : sachant que les données sur les réseaux d'assainissements sont issues de plusieurs sources comment pourrons-nous les fusionner pour avoir des réseaux plus certains, précis et complets ?

La fusion des données sur les réseaux d'assainissement a pour but d'enrichir une base de données avec des données provenant d'autres sources. Par exemple, une base de données contient une géométrie précise mais il lui manque de nombreux attributs. Ceci permet de créer

une nouvelle base de données complète à partir de plusieurs sources de données incomplètes et de détecter les incohérences dans les données disponibles de manière à améliorer les mesures de confiance des données disponibles.

1.3 Contribution

Notre contribution concerne principalement :

- L'étude des imperfections des données rencontrées lors de la fusion des données des réseaux d'assainissement multi-source.
- La catégorisation des imperfection liées aux nœuds et aux arêtes des réseaux d'assainissement.
- L'adaptation du processus d'appariement géographique proposée par [29] à notre contexte en utilisant la théorie des fonctions de croyance, pour permette la gestion des imperfections de chaque catégorie lors de la fusion des données des réseaux d'assainissement.
- La modélisation sémantique des critères de similarité pour fusionner les données des réseaux d'assainissement multi-source.

Chapitre 2

Théorie des fonctions de croyance

2.1 Introduction

Dans le cadre des systèmes d'aide à la décision, l'analyse des données et la reconnaissance des formes, nous manipulons des informations qui sont de plus en plus nombreuses dans la plupart des domaines, qui sont normalement censées donner une image aussi juste que possible de la réalité. Or, le plus souvent, ces informations sont imparfaites (voir l'Annexe A.1).

Pour modéliser ces imperfections, nous faisons appel à la théorie des fonctions de croyance, aussi appelée dans la littérature théorie de Dempster-Shafer et plus particulièrement théorie de l'évidence [1], la théorie des fonctions de croyance est issue des travaux de Dempster [11] sur la généralisation de l'inférence bayésienne quand il n'y a pas de connaissance a priori sur les paramètres. Cette généralisation conduit à l'utilisation de probabilités non-additives qui peuvent être combinées par une règle. Cependant, le développement du formalisme de la théorie est attribué à Shafer [40].

En 1976, Shafer a montré l'intérêt des fonctions de croyance pour la modélisation des connaissances incertaines. Elles permettent également de représenter l'information imparfaite d'une façon plus naturelle qu'avec des probabilités. En effet, la représentation de ces informations imparfaites, ou l'absence d'information, n'est pas bien prise en considération par la théorie des probabilités. Dans ce chapitre, nous détaillons les notions fondamentales de la théorie des fonctions de croyance à travers les diverses fonctions développées. De plus nous présentons les méthodes les plus connues pour estimer les fonctions de masse.

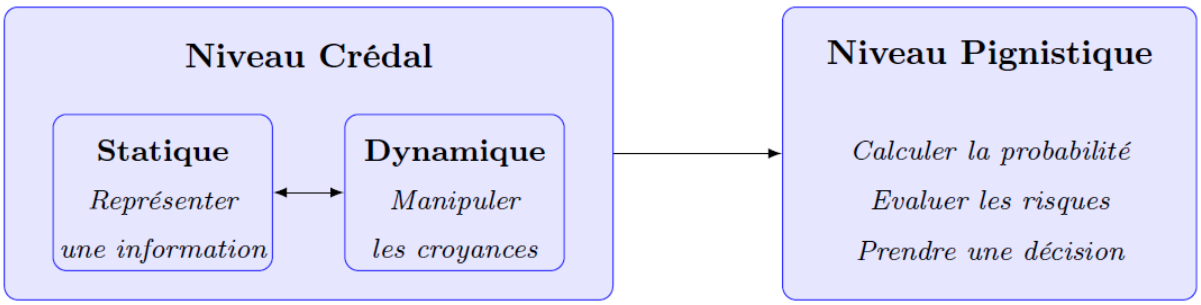


FIGURE 2.1 – Schéma des deux niveaux du Modèle des Croyances Transférables MCT [1]

2.2 Niveau crédal statique : Représentation des informations

Dans ce qui suit, nous présentons les différentes opérations de base réalisées au niveau crédal statique.

2.2.1 Cadre de discernement

La théorie des fonctions de croyance est fondée sur plusieurs notions, parmi lesquelles on trouve le cadre de discernement ou cadre d'intérêt noté Ω. l'ensemble Ω désigne l'ensemble des réponses possibles (ω_i) à un problème :

$$\Omega = \{w_1, \ldots, w_K\} \tag{2.1}$$

Ce cadre peut valider la notion de monde fermé [45], autrement dit, l'ensemble est exhaustif. Il est également possible de se passer de cette condition en prenant Ω en tant que cadre non exhaustif. Cette démarche est alors appelée l'hypothèse du monde ouvert [46]. A partir du cadre de discernement Ω, on déduit l'ensemble 2^Ω, comprenant l'ensemble des 2^Ω sous-ensembles $A \subseteq \Omega$:

$$2^\Omega = \{A, A \subseteq \Omega\} = \{\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_K\}, \{\omega_1 \cup \omega_2\}, \ldots, \Omega\} \tag{2.2}$$

Cet ensemble regroupe les différentes hypothèses de Ω mais également toutes les disjonctions possibles de ces hypothèses. Cet ensemble permet de définir l'ensemble des quantités utilisées par la théorie des fonctions de croyance afin d'évaluer la vérité d'une proposition.

2.2.2 Fonction de masse

Soit une question Q à laquelle on cherche une réponse, soit le cadre de discernement contenant les solutions possibles à cette question $\Omega = \{w_1, \ldots, w_K\}$. Une fonction m, appelée *fonction de masse* (en anglais basic belief assignment ou bba) représente une connaissance imparfaite sur Ω, est définie de 2^Ω à valeurs dans l'intervalles [0,1], avec comme contrainte :

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{2.3}$$

La quantité $m(A)$ représente la croyance sur le fait que A contienne la réponse à la

question Q, un ensemble A tel que $m(A) > 0$ est appelé un élément focal, et la réunion des éléments focaux constituent le noyau. À l'origine Shafer [40] a ajouté une contrainte supplémentaire $m(\phi) = 0$, ce qui signifie l'exhaustivité de l'ensemble Ω et qui fait référence au monde fermé, dans le MCT cette contrainte est relâchée ce qui fait référence au monde ouvert, la quantité $m(\phi)$ peut-être interprétée comme la masse de croyance dans le fait que la réponse de Q se trouve ailleurs de Ω.

2.2.3 Fonctions de masse particulaires

La section suivante contient les définitions de certaines fonctions de masse particulaires.

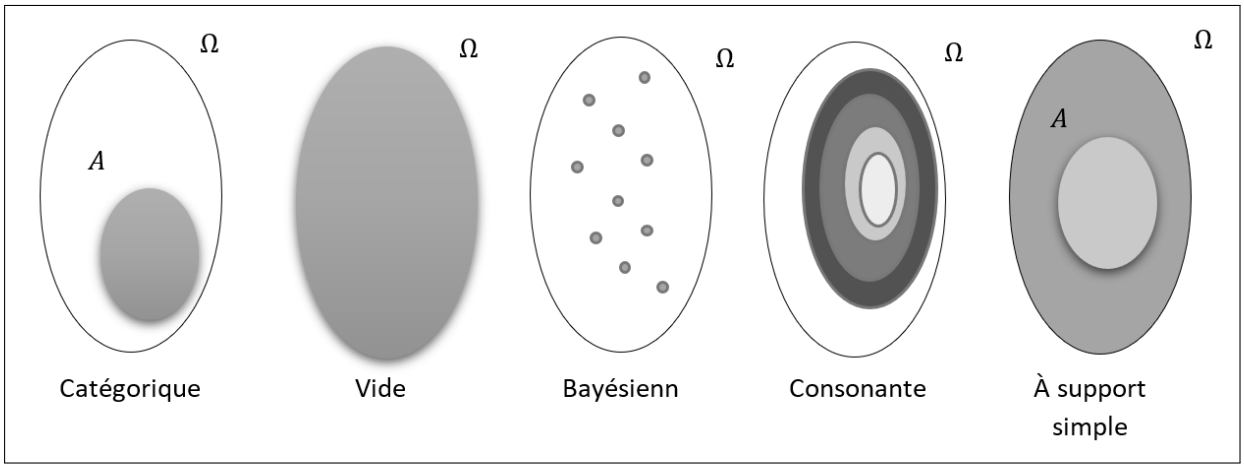


FIGURE 2.2 – Fonctions de masses particulaires

- Fonction de masse catégorique sur $A \subseteq \Omega$ si $m(A) = 1$.
- Fonction de masse vide ou d'ignorance totale lorsque la fonction m est catégorique en $\Omega : m(\Omega) = 1$.
- Fonction de masse Bayésienne si tous les éléments focaux sont des singletons de Ω.
- Fonction de masse consonante si tous les éléments focaux sont emboîtés.
- Fonction de masse à support simple s'il possède deux éléments focaux dont l'un est Ω.
- Fonction de masse normale si $m(\Omega) = 0$.

2.2.4 Autres fonctions de croyance

Les croyances peuvent être allouées à d'autres fonctions qui sont liées à la fonction de masse m :

- La fonction de croyance $\mathrm{bel}(A)$, aussi appelée crédibilité, qui permet de quantifier la croyance totale en A est définie par :

$$\mathrm{bel}(A) = \sum_{\phi \neq B \subseteq \Omega} m(B) \quad \forall A \subseteq \Omega \tag{2.4}$$

  Par exemple, dans la figure 2.2.4 $\mathrm{bel}(B_i) = A_1 + A_3$.

- La fonction de plausibilité définie par :

$$\begin{aligned} pl(A) &= \sum_{A \cap B} m(B) \quad \forall A \subseteq \Omega \\ &= \mathrm{bel}(\Omega) - \mathrm{bel}(\bar{A}) \\ &= 1 - m(\emptyset) - \mathrm{bel}(\bar{A}) \end{aligned} \tag{2.5}$$

  La quantité $pl(A)$ représente la part de croyance qui pourrait potentiellement être allouée à A. Par exemple, dans la figure 2.2.4 $\mathrm{pl}(B_i) = A_1 + A_2 + A_3 + A_4$.

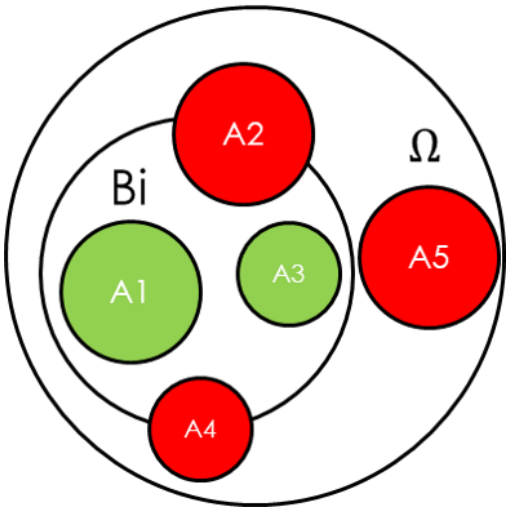


FIGURE 2.3 – Exemple de crédibilité

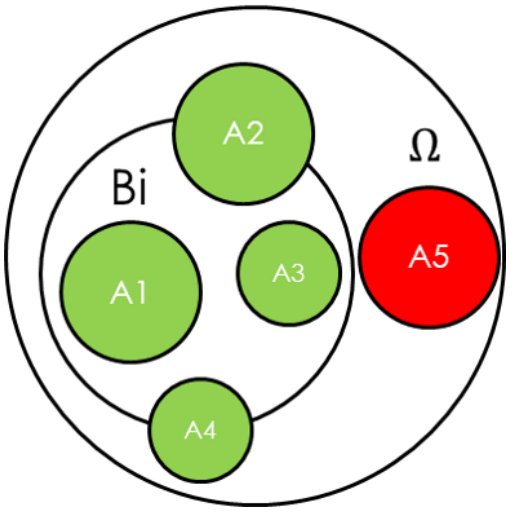


FIGURE 2.4 – Exemple de plausibilité

- La fonction de communalité est définie par :

$$q(A) = \sum_{B \supseteq A} m(B) \quad \forall A \subseteq \Omega \tag{2.6}$$

La quantité $q(A)$ s'interprète comme la totalité de la croyance accordée aux sur-ensembles de A.

- La fonction d'implicabilité est définie par :

$$b(A) = \sum_{A \subseteq B} m(B) \quad A \subseteq \Omega \tag{2.7}$$

La quantité $b(A)$ s'interprète comme la totalité de la croyance accordée aux sous-ensembles de A.

2.2.5 Opérations sur la représentation des connaissances : raffinement / grossissement

La construction du cadre de discernement est une étape très intéressante. Cependant, il est généralement difficile de définir précisément la granularité du cadre dès le départ. Il existe donc des opérations qui permettent de modifier la granularité du cadre de discernement. Soit deux cadres de discernement Θ et Ω, la fonction ρ, une opération de 2^{Ω} dans 2^{Ω}. La figure 2.5 présente un grossissement, défini par $\rho(\{\theta_1\}) = \{W_1, W_2\}$ et $\rho(\{\theta_2\}) = \{W_3, W_4\}$.

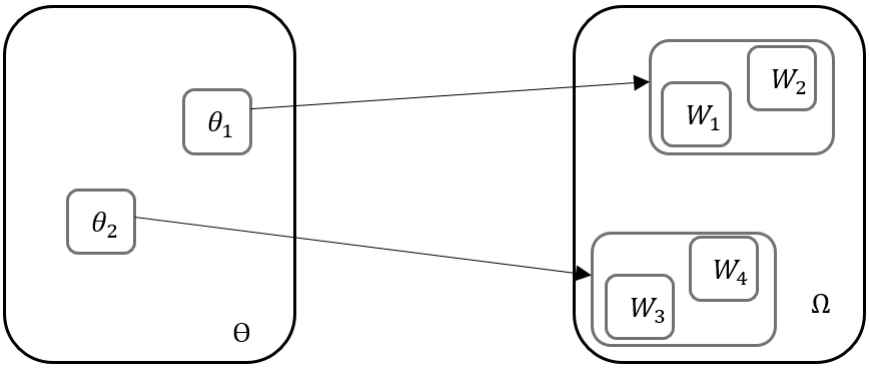


FIGURE 2.5 – Grossissement / Raffinement

2.3 Niveau crédal dynamique : manipulation des informations

Dans cette section, nous présentons un certain nombre de concepts de la théorie des fonctions de croyance qui rendent possible le transfert de croyances entre les sous-ensembles de Ω.

2.3.1 Conditionnement et déconditionnement

Soit une masse $m(A)$ avec $A \subseteq \Omega$. S'il y a une information supplémentaire sûre et fiable qui affirme que la véracité existe dans un ensemble $B \subseteq \Omega$, alors la masse initialement allouée à A est transférée à $A \cap B$. Cette opération est appelée le conditionnement. La fonction de masse m conditionnée par B s'écrit $m[B](C)$ et est définie par :

$$m[B](C) = \begin{cases} \sum_{A \cap B = C} m(A), & \text{si } C \subseteq B \\ 0, & \text{sinon} \end{cases} \tag{2.8}$$

Néanmoins, la masse allouée à $m[B](\phi)$ qui représente la croyance de m incompatible avec l'information certifiant que la vérité se trouve dans B peut-être non nulle. Cependant, nous ajoutons une étape de normalisation :

$$m^*[B](C) = \begin{cases} \frac{m[B](C)}{1 - m[B](\phi)}, & \text{si } A \subseteq \Omega, \text{A} \neq \phi \\ 0, & \text{A} = \phi \end{cases} \tag{2.9}$$

L'opération inverse du conditionnement est le déconditionnement. Par exemple, soit une masse m[B] conditionnelle à $B \subseteq \Omega$, de manière générale, il existe plusieurs fonctions de croyance m qui mènent à la masse conditionnée $m[B]$. En appliquant le Principe du Minimum d'Information (PMI) [42], qui permet de sélectionner la fonction de masse la moins engageante, c'est-à-dire que les masses doivent être transférées sur les sous-ensembles les plus grands, la fonction de masse déconditionnée $m[B]^{\uparrow}$ est définie par :

$$\begin{cases} m[B]^{\uparrow}(A \cup \bar{B}) & = m[B](A) \quad \forall A \subseteq B \\ m[B]^{\uparrow}(C) & = 0 \quad \forall C \subset \bar{B}, C \neq \emptyset \end{cases} \tag{2.10}$$

2.3.2 Affaiblissement

Quand une connaissance a priori concernant la confiance d'une source S qui a donné une fonction de masse m est présente, il est possible de prendre en compte cette méta-connaissance afin d'effectuer une opération d'affaiblissement, soit $\alpha \in [0, 1]$, un coefficient d'affaiblissement, la fonction de masse affaiblie est obtenue par :

$$\begin{cases} \alpha_{m(\text{ A})} = (1-\alpha)m(\text{ A}), \quad \forall A \subset \Omega \\ \alpha_{m(\Omega)} = (1-\alpha)m(\Omega) + \alpha \end{cases} \tag{2.11}$$

Le degré de fiabilité de la source S est $(1-\alpha)$. Dans le cas où $(1-\alpha) = 0$, la source S n'est pas du tout fiable et la totalité de la masse est reportée à Ω, Cela donne une fonction de masse vide (ignorance totale). Par contre, lorsque la source est absolument fiable $(1-\alpha) = 1$, la fonction de masse m n'est pas affaiblie.

2.3.3 Combinaison de fonction de croyance

Dans le contexte de données imparfaites (incertaines, imprécises et incomplètes), la combinaison de données est une bonne solution pour obtenir des informations plus significatives. La phase de combinaison, dans le contexte de la théorie des fonctions de croyance, permet donc de synthétiser, sous la forme d'une unique fonction de croyance, l'ensemble des connaissances issues de plusieurs fonctions. Cette étape de fusion a pour objectif d'exploiter la complémentarité et la redondance des différentes sources afin d'obtenir une connaissance générale sous la forme d'une fonction de croyance plus précise et plus renforcée.

La méthode de combinaison de base, présentée dans le cadre de la MCT, est la règle de combinaison conjonctive. Pour deux fonctions de masses $m1$ et $m2$, cette règle est définie par :

$$m_1 \oplus_2 (A) = m_1 \oplus m_2(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \quad \forall A \subseteq \Omega \tag{2.12}$$

La fonction de masse m correspond donc à la synthétisation des connaissances de m1 et m2. Dans le cas de J fonctions à combiner, la formule de combinaison s'écrit :

$$m_{\oplus}(A) = \oplus_{j=1}^{J} m_j(A) = \frac{1}{1-K} \sum_{A_1 \cap \ldots \cap A_j = A} \left(\prod_{j=1}^{J} m_j\left(A_j\right) \right) \quad \forall A \subseteq \Theta, A \neq \emptyset \tag{2.13}$$

Cette règle de combinaison conjonctive permet de vérifier plusieurs propriétés comme l'associativité, la commutativité et dispose d'un élément neutre qui est la fonction de masse vide.

L'une des particularités de cette règle est de générer une fonction de masse non normalisée $m(\emptyset) > 0$. Cette condition n'étant pas possible dans un monde fermé, une phase de normalisation est nécessaire.

La normalisation de la combinaison conjonctive est appelée la règle de combinaison orthogonale ou la règle de combinaison de Dempster. Elle est également la première règle présentée par la théorie et s'écrit comme suit :

$$m_{1\oplus 2}(A) = m_1 \oplus m_2(A) = \frac{1}{1-\kappa} \sum_{B\cap C=A} m_1(B)m_2(C) \quad \forall A \subseteq \Omega, A \neq \emptyset \tag{2.14}$$

où :

$$\kappa = m_{1\oplus 2}(\emptyset) = \sum_{B\cap C=\emptyset} m_1(B)m_2(C). \tag{2.15}$$

κ est la masse de conflit qui prend des valeurs entre 0 et 1. En général, elle traduit le degré de contradiction entre les sources combinées. Pour $\kappa = 1$, $S1$ et $S2$ sont entièrement considérées comme conflictuelles et les sources ne sont pas fusionnées. Au contraire, si $\kappa = 0$, les sources sont parfaitement en accord.

Les deux règles de combinaison (conjonctive et de Dempster) ne sont pas idempotentes. Par conséquent, elles ne peuvent pas être appliquées lorsque l'indépendance des sources n'est pas vérifiée. La définition de la notion d'indépendance est complexe, bien que de nombreux travaux aient été effectués sur cet objectif dans les fonctions de croyance [12, 18, 55]. Cependant, l'indépendance cognitive proposée par Smets [42] est fréquemment utilisée. Dans ce cas, les sources sont considérées comme indépendantes quand il n'y a pas eu de consultation entre elles avant l'émission des croyances.

En général, les opérateurs conjonctifs sont destinés à la combinaison de fonctions de masse issues de sources d'information réputées fiables. Lorsqu'au moins une des fonctions de croyance à combiner n'est pas fiable, La première règle élaborée est la règle de combinaison disjonctive. Pour deux fonctions de croyance m_1 et m_2, cette règle est définie par :

$$m_{1\cup 2}(A) = m_1 \cup m_2(A) = \sum_{B\cup C=A} m_1(B)m_2(C) \quad \forall A \subseteq \Omega \tag{2.16}$$

Cette combinaison est la règle inverse de la combinaison conjonctive. Elle est aussi associative et commutative mais non-idempotente. Dans la pratique, cette règle est peu utilisée car elle élargit les éléments focaux et donc perd en spécificité, ce qui peut compliquer la décision finale. De plus, elle est abordée dans le cadre du théorème de Bayes généralisé de Smets [42].

2.4 Niveau pignistique : Prise de décision

La manipulation de l'information est propre à la partie crédale où une synthèse complète est produite sous la forme d'une fonction de croyance m. Pour ce qui est de la phase décisionnelle, elle concerne le niveau pignistique. La phase de prise de décision est basée sur la répartition pignistique [43] notée BetP et obtenue à partir de la fonction de masse m. Elle est aussi nommée probabilité pignistique pour l'ensemble des probabilités sur les singletons qu'elle génère. Le transfert de la fonction de masse m en une fonction de probabilité pignistique BetP exprimée sur Ω, se caractérise pour tout $\omega_k \in \Omega$ par :

$$\mathrm{BetP}\left(\omega_k\right) = \frac{1}{1 - m(\emptyset)} \sum_{A \ni \omega_k} \frac{m(A)}{|A|} \tag{2.17}$$

Après cette transformation, il est impossible de retrouver la fonction de masse initiale m. En effet, une fonction de masse n'est liée qu'à une unique probabilité pignistique. Par contre, une probabilité pignistique peut s'obtenir à partir d'un nombre illimité de fonctions de masse. Généralement, la décision se fait en choisissant l'élément ω_k dont la probabilité pignistique est la plus élevée.

2.5 Gestion du confit

La fusion de données est un moyen très intéressant d'obtenir des informations plus pertinentes. Historiquement, l'opérateur de Dempster est le premier opérateur de combinaison défini dans la théorie des fonctions de croyance. Cette règle permet de vérifier quelques propriétés importantes et son usage a été théoriquement justifié par plusieurs auteurs [15, 21]. Cependant, son application n'est pas justifiable dans tous les contextes et il existe des cas où son utilisation peut conduire à des résultats incohérents. En effet, nous pouvons distinguer des situations où les sources d'information utilisées sont dépendantes ou manquent de fiabi-

lité. C'est Zadeh [60, 58] qui a montré, à l'aide d'un exemple, l'incohérence du résultat de la combinaison orthogonale dans le contexte de sources non fiables. Dans ces cas, il est fortement déconseillé d'utiliser la somme orthogonale. Lors de la combinaison conjonctive, il peut apparaître une masse vide $m(\emptyset)$ plus ou moins importante selon la contradiction entre les sources d'information combinées.

Diverses règles de combinaison sont proposées pour régler ce conflit. Nous distinguons deux types d'approches. La première consiste à agir sur les sources avant leur fusion en détectant les sources non fiables. La source non fiable est alors affaiblie. De nombreux travaux se sont intéressés à l'estimation de la fiabilité des sources d'information à partir de mesures de distance. Dans ce cadre, on peut citer les travaux [56, 19, 25, 38, 37]. La deuxième solution, qui a pour but de répartir le conflit plus finement, consiste à le traiter au niveau de la combinaison. La liste des opérateurs de cette famille ne se limite pas à la combinaison orthogonale mais également à d'autres approches [41].

2.6 Conclusion

Nous avons exploré dans ce chapitre un certain nombre de principes fondamentaux de la théorie des fonctions de croyance. La problématique de la représentation des fonctions de croyance a donné lieu à plusieurs interprétations. Parmi ces interprétations, nous trouvons le Modèle des Croyances Transférables qui est une interprétation non probabiliste. Il s'agit également d'une lecture multi-niveau de la théorie où deux niveaux peuvent être distingués. Dans le premier niveau, les croyances sont manipulées et deux sous-niveaux peuvent être distingués. Dans le niveau crédale statique, l'information est modélisée et dans le niveau crédale dynamique, les croyances sont manipulées. Enfin, dans le deuxième niveau, le niveau pignistique, la décision est prise par le biais de la probabilité pignistique.

Nous présentons dans l'Annexe B un aperçu des autres théories faisant partie de la même famille des théories de l'incertain : la théorie des probabilités, la théorie des ensembles approximatifs et la théorie des possibilités.

Chapitre 3

Appariement des objets géographiques

3.1 Définition

Dans la littérature, plusieurs définitions ont été proposées. Dans [50] l'appariement des objets est l'identification d'objets correspondants dans différentes sources de données. Dans [49] il est utilisé pour déterminer les objets correspondants dans différents ensembles de données qui représentent des entités analogues dans le monde réel. Dans [27] est défini comme la détection des objets homologues de deux jeux de données. L'interprétation de ces définitions est identique. Le mot 'objets' fait référence à des points, des lignes ou des Polygones.

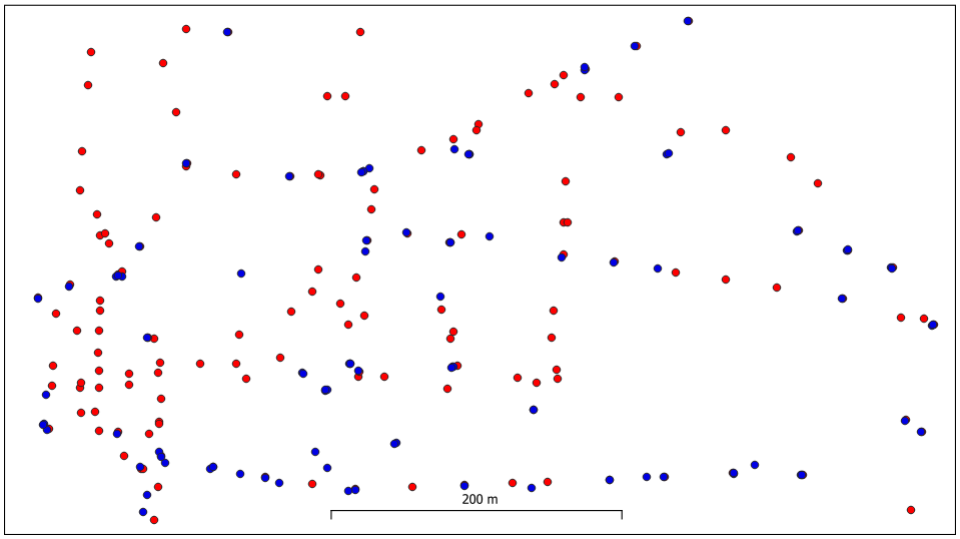


FIGURE 3.1 – Visualisation de deux source de données spatiales

Supposons que nous souhaitions matcher les objets superposés de la figure 3.1, sur laquelle nous avons deux sources de données. La première est représentée par des points rouges l'autre par des points bleus. Les premières questions qui viendront probablement à l'esprit :

- Dans quelles mesures nous pouvons juger que deux objets sont similaires ?
- Y a-t-il un/des moyen(s) pour mesurer la similarité entre deux objets spatiaux ? Si oui, ces moyens sont-ils nombreux ?
- Comment savoir si un objet n'a pas d'homologue ?
- Comment utiliser ces mesures pour identifier des homologues ?

Généralement, pour effectuer l'appariement d'objets nous avons besoin de deux ingrédients (figure 3.2) :

- Une ou plusieurs mesures de similarités.
- Une méthode qui utilise ces mesures pour Matcher les objets.

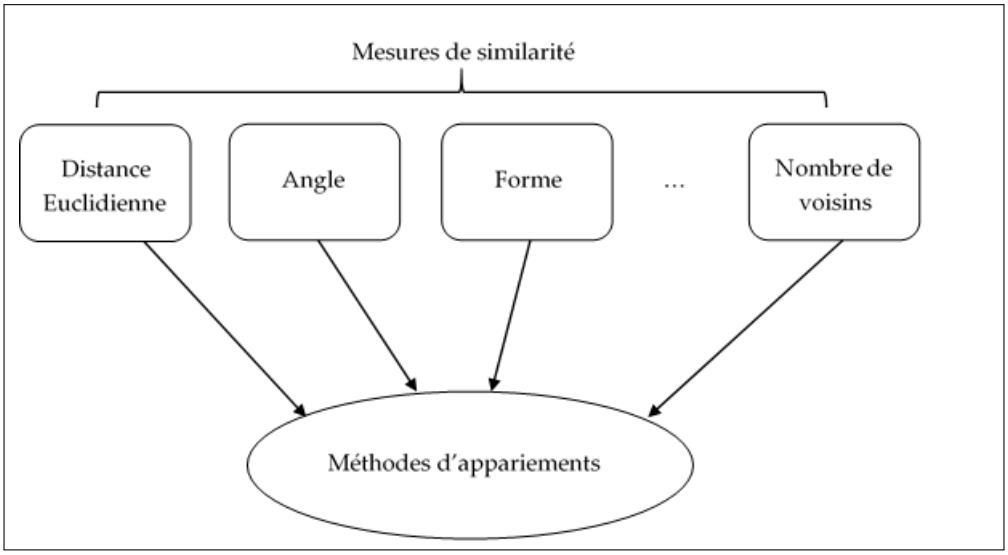


FIGURE 3.2 – Ingrédients d'appariement d'objets.

3.2 Les mesures de similarité

Les mesures de similarité sont utilisées pour obtenir un certain degré de comparaison entre les instances [32]. Elles représentent les critères sur lesquels se base un appariement [23]. La mesure de similarité la plus intuitive est celle qui s'appuie sur la position, de telle sorte que nous supposons que deux objets sont homologues, lorsqu'ils sont proches en termes de distance. La plus utilisée est la distance l'euclidienne [3, 36, 50]. 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, nous pouvons penser à utiliser des attributs, le nombre de voisins, la topologie etc.

Historiquement, selon [52, 53, 47], la majorité des premières approches 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, 35, 34]. Depuis, un grand nombre de mesures ont été proposées. Nous les classifions en 3 catégories :

- Géométrique
- Topologique

- Attributaire

3.2.1 Mesures Géométrique

3.2.1.1 Position

la position géographique n'est pas en elle-même la mesure, mais c'est sur sa base que nous pouvons calculer la distance entre les objets. Il existe plusieurs distances. Dans [36, 50, 52, 23, 49, 28] ils utilisent la distance euclidienne. Aussi dans [3], en justifiant le choix de la position comme la seule propriété toujours disponible dans les données spatiales, ils exploitent la distance euclidienne pour calculer la probabilité que le i -ème objet de la source une a_i est le correspondent du j -ème objet de la source deux b_i :

$$\mathrm{P}_{a_i}(b_j) = \frac{\text{distance}\,(a_i, b_j)^{-\alpha}}{\sum_{k=1}^{m} \text{distance}\,(a_i, b_k)^{-\alpha}} \tag{3.1}$$

où m est le nombre d'objets de la source deux et α un paramètre à fixer, qui indique combien la valeur de la probabilité va décroître quand la distance entre les objets augmente.

Dans [50, 24] la distance Hausdorff a été utilisée sur des lignes (figure 3.3). Cette distance est définie comme la distance maximale d'un ensemble par rapport à l'objet le plus proche dans l'ensemble cible. Dans [13, 63, 20] les auteurs ont utilisé la distance comme filtre pour éliminer les candidats qui dépassent une valeur maximale. Par exemple, deux routes ne peuvent pas être appariées si la distance est supérieure à 15 mètres.

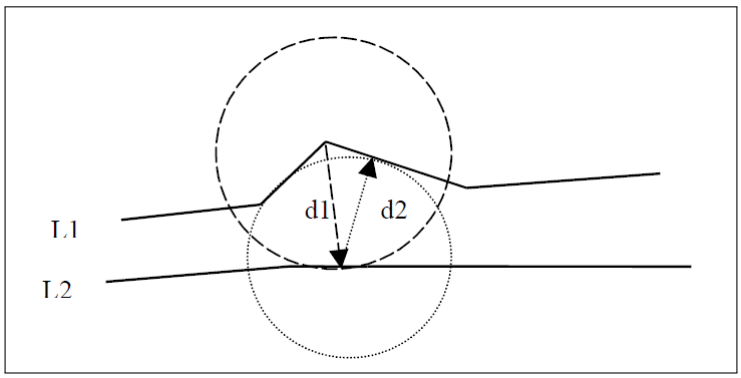


FIGURE 3.3 – La distance Hausedouff entre deux lignes L1 et L2 [57]

3.2.1.2 Forme

Dans [51], la forme est classée en trois catégories en fonction de la distance maximale des sommets par rapport à une ligne imaginaire reliant le point de départ au point d'arrivée (voir figure 3.4). Il s'agit d'une mesure très simple pour déterminer la forme d'une ligne.

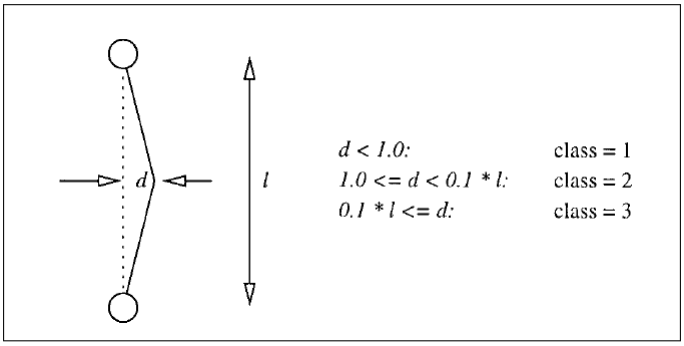


FIGURE 3.4 – Calcul de la forme

La figure 3.5 montre les probabilités qu'un élément appartienne à une classe de forme spécifique et la figure 3.6 montre les probabilités conditionnelles. La situation la plus probable est qu'un élément est apparié à un élément qui est dans la même classe. Plus la distance entre deux classes est grande, plus l'appariement est improbable.

	$P_{form}(a_i)$
$a_i \in$ class 1	0.59
$a_i \in$ class 2	0.27
$a_i \in$ class 3	0.14

FIGURE 3.5 – Distribution de probabilité de la forme

$P_{form}(a_i \mid b_j)$	$b_j \in$ class 1	$b_j \in$ class 2	$b_j \in$ class 3
$a_i \in$ class 1	0.87	0.11	0.02
$a_i \in$ class 2	0.28	0.61	0.11
$a_i \in$ class 3	0.09	0.14	0.77

FIGURE 3.6 – Probabilités conditionnelles de la forme

3.2.2 Mesures Topologiques

3.2.2.1 Degré d'un noeud

Il s'agit de l'une des mesures topologiques les plus utilisées. Elle représente le nombre de lignes qui sont connectées à un noeud (figure 3.7). Nous retrouvons cette mesure dans plusieurs publications dont [32, 63, 47].

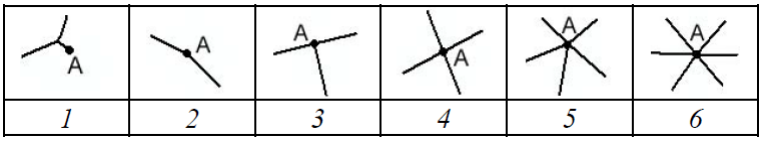


FIGURE 3.7 – Degré d'un noeud [63]

3.2.2.2 Direction des lignes

Dans [32] cette mesure représente la direction des lignes émanant d'un noeud. Puisqu'elle est une valeur continue (angle), et pour qu'elle ne soit pas trop sensible aux petits changements, les auteurs ont défini 8 intervalles basés sur des axes virtuels, tels que deux lignes ont la même orientation si elles appartiennent au même intervalle (figure 3.8).

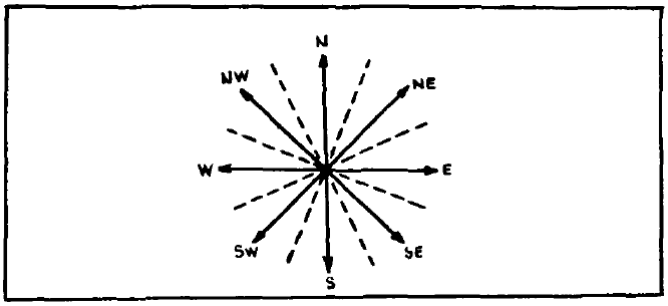


FIGURE 3.8 – Discrétisation de l'orientation des lignes selon 8 axes [32]

Dans [30], l'orientation est calculée entre les lignes de deux sources différentes. Elle est évaluée en calculant la différence d'angle entre deux lignes par rapport à la tangente, telle que, si l'angle θ est proche de 0, les objets ont une orientation similaire, si $\theta = \pi$ ils sont dans le sens inverse et si $\theta = \pi/2$ alors ils sont perpendiculaires.

3.2.2.3 Autres

Dans [4], une présentation d'un grand nombre de relations topologiques que nous pouvons exploiter (figure 3.9). Dans [30], ils ont utilisé le rapport des surfaces qui se chevauchent pour apparier des Polygones. Dans [36], en partant de l'hypothèse que le contexte géographique des objets est invariant et indépendant de la représentation, ils proposent une nouvelle mesure de similarité nommée graphe de proximité qui consiste à dessiner un graphe entre l'objet à apparier et ces adjacents, de même pour les candidats possibles et par la suite une comparaison de ces graphes en matière de distance et d'angle est effectuée pour juger si deux objets se correspondent (figure 3.10). [20] s'inspirent de ce travail pour proposer une méthode

qui utilise le diagramme de Voronoi et la triangulation de Delaunay pour capturer le contexte.

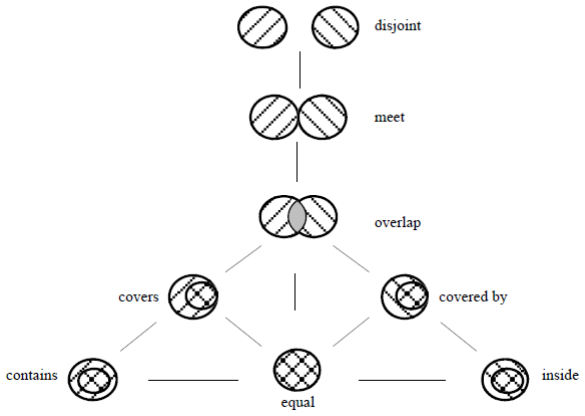


FIGURE 3.9 – relations topologiques des voisins [4]

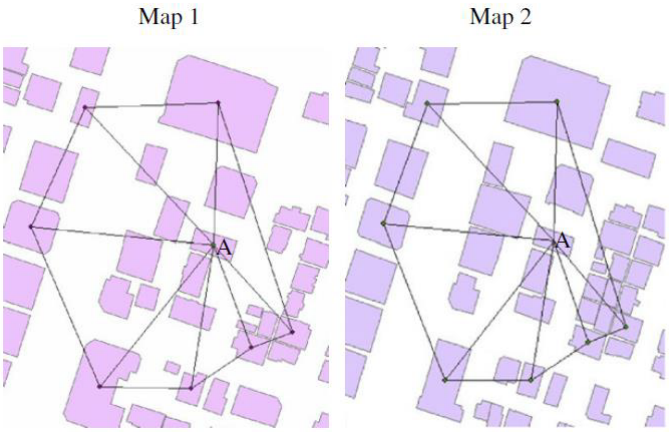


FIGURE 3.10 – Graphe de proximité [20]

La majorité des mesures de similarités que nous avons présentées dans le cadre topologique, exploitent les relations entre les objets du même type appartenant au même ensemble de données. Par exemple, dans le graphe de proximité sur la figure 3.10, le graphe du bâtiment 'A' qui capture ces relations spatiales est construit en utilisant les bâtiments voisins.

Cependant, dans ce travail [9] qui vise à apparier les réseaux hydrauliques, au lieu d'utiliser les voisins du même ensemble, c'est-à-dire les arêtes du réseau pour capturer les relations spatiales, d'autres objets appartenant à un autre ensemble de données sont exploités. La figure 3.11 illustre le critère utilisé pour comparer deux bases de données (Cassini et BDCarto) qui sont composées de lignes (réseaux hydrauliques) et points (des localités, villages, moulins etc.). Sachant que dans un travail précèdent les auteurs ont apparié les points (matching : ligne en

noir), ils partent de l'hypothèse que si deux lignes de réseaux hydrauliques se correspondent, les points qui les entourent probablement doivent aussi s'apparier. Ils utilisent alors ce nouveau critère pour évaluer la similitude des lignes, à travers le calcul du ratio des points qui matchent et qui sont proches des traits candidats. Nous signalons que cela n'était pas possible sans les résultats de l'appariement à priori des points, mais cela reste une idée nouvelle par rapport aux autres critères de similitudes.

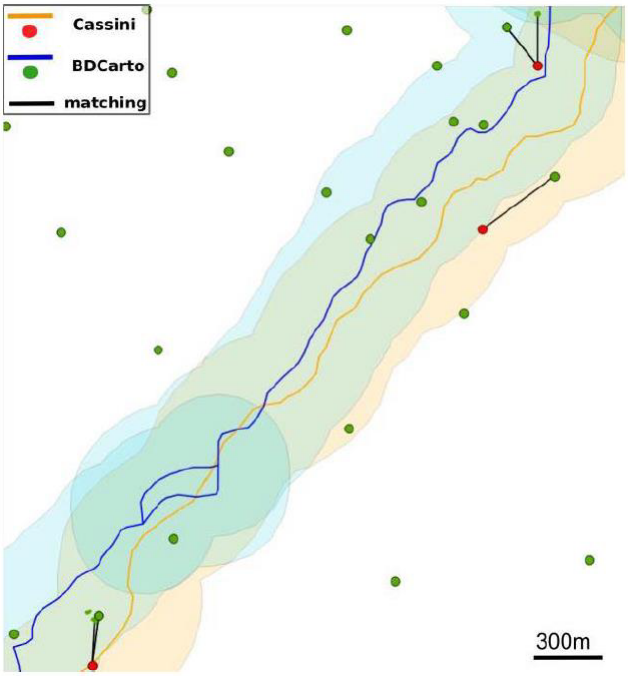


FIGURE 3.11 – Introduction d'un nouveau critère basé sur la distance aux entités ponctuelles déjà appariées [9]

3.2.3 Mesures Attributaires

Elles consistent à comparer les attributs associés aux objets spatiaux, comme les noms et les adresses. La plus utilisée est la distance de Leiveishtein qui est définie comme le nombre minimum d'insertions, suppressions ou substitutions nécessaires pour transformer une chaîne de caractères à une autre [4, 36, 9].

Parfois, la sémantique des attributs est utilisée pour réduire l'ensemble des objets avec lequelle il est possible de faire l'appariement. Il est possible d'apparier des objets qui sont de type 'peak' et 'Summit' mais pas 'river' et 'summit' [30]. Pour appliquer ce raisonnement et identifier cette différence automatiquement, ils font appel aux ontologies. Aussi, dans [20] la valeur des attributs permet d'identifier les champs qui sont identiques. Cependant, nous jugeons que l'utilisation de ces approches pour identifier les champs attributaires fait partie de

l'appariement des schémas et non pas l'appariement des objets. Car, lorsque nous sommes en train de comparer la sémantique des attributs nous nous sommes plus au niveau des instances mais plutôt au niveau des modèles de données et la relation de leurs éléments.

Nous avons présenté dans les paragraphes précédents un ensemble de mesures de similarité classifiées selon trois catégories : géométriques, topologiques et attributaires. Dans la littérature, il existe plusieurs classifications. Dans [32], ils présentent les mesures sous les catégories suivantes : géométriques/topologiques, dépendant/indépendant et continue/discret. Dans [36], ils les séparent en deux classes. Contexte-indépendant : ce sont des mesures basées sur les attributs individuels des objets sans égard à leurs relations avec les autres objets, comme la longueur et les attributs. Contexte-dépendant : fait référence aux relations spatiales entre les objets d'une zone. Dans [30], quatre types sont définis : géométriques, topologiques, attributaire et sémantiques. Ici, le mot 'sémantique' fait référence aux types d'objets, et comme nous l'avons indiqué précédemment cela relève de l'appariement des schémas. Dans [54], cinq catégories : géométriques, topologique, attributs, contexte et sémantique. La même remarque précédente pour le type sémantique. Quant au contexte, nous jugeons qu'il fait partie de la topologie vu qu'il fait appel aux relations spatiales entre les objets, c'est pourquoi dans notre classification nous l'avons inclus dans le type topologique (le graphe de proximité). Nous retrouvons la classification qui a été adoptée dans plusieurs publications dont [57, 47]. Cependant, la signification de chaque classe est différente d'un travail à un autre. Par exemple, dans [57] la classe attribut inclus leurs sémantiques.

Pour conclure cette partie, nous rappelons l'essentiel. Les mesures de similarités sont un moyen pour quantifier la similitude entre des objets appartenant à des ensembles de données différents. Les objets peuvent être des points, des lignes ou des polygones. Dans la littérature il en existent plusieurs, les plus utilisées sont les géométriques, car les informations nécessaires pour les utiliser sont généralement disponibles : positions et formes. Les autres classes, topologiques et attributaires sont employées souvent comme complémentaires aux mesures géographiques. La question que nous pouvons nous poser maintenant est : comment utiliser ces mesures pour apparier les objets ? et comment coordonner l'utilisation de ses différents critères ? Les réponses à ces questions seront présentées dans la section suivante.

3.3 Les méthodes d'appariements

3.3.1 Classification selon le nombre de mesures de similarités

Comme nous l'avons précisé dans la section précédente, ces méthodes utilisent les mesures de similarités pour identifier l'appariement. Sur cette base nous pouvons faire la distinguions entre deux catégories :

- **Monocritères** : utilisent une seule mesure pour matcher les objets. [3], appliquent seulement le critère de distance pour apparier les objets.
- **Multicritères** : La catégorie la plus populaire. Les méthodes utilisent plusieurs mesures pour apparier les objets. [36, 28, 9], appliquent un ensemble de critères géométriques, topologiques et attributaires. Plusieurs publications associent des poids à ses critères pour favoriser les uns sur les autres et en extraire une mesure de similarité globale. Nous pouvons classifier le choix de ces poids en deux catégories :
  - **Par un expert**, comme dans [36, 9].
  - **Par apprentissage**, comme dans [63, 52].

Nous pouvons aussi classifier les méthodes selon le type des mesures qu'elles utilisent : géométriques ou/et topologiques ou/et attributaires.

3.3.2 Classification selon la relation des méthodes avec les mesures

La relation des méthodes et les mesures de similarité peuvent varier d'une approche à une autre. Nous avons distingué deux types :

- **Générique** : les méthodes qui prennent les mesures comme paramètres, c'est-à-dire l'utilisateur peut adapter l'approche à sa problématique en modifiant seulement les mesures de similarités. Comme dans [24] qui transforment l'appariement en problème d'optimisation.
- **Ad-Hoc** : Des méthodes spécifiques à une problématique, dont les mesures font partie intégrante. Par conséquent, ces méthodes ne sont pas réutilisables. Comme dans [9] où le critère de similarité est basé sur la structure des réseaux hydrauliques.

3.3.3 Classification selon la direction d'appariement

Supposons que nous avons identifié le type d'application. Par exemple, point, échelle similaire et que les cardinalités ciblent 1 :1, et que nous souhaitons parcourir nos objets pour identifier ceux qui se correspondent, deux stratégies se présentent :

- **Monodirectionnelle** : Parcourir les objets d'un ensemble dit de référence puis chercher pour chacun un correspondant dans les objets appartenant à(aux) l'ensemble(s) cible(s). Il existe des solutions comme l'algorithme de plus proche voisin qui impose par définition à ce que chaque objet de l'ensemble de référence ait un correspondant ce qui n'est pas toujours le cas en réalité.
- **Bidirectionnelle** : Chacun des ensembles à intégrer est utilisé comme ensemble de référence et cible [52, 3]. Par exemple, dans le cas de deux sources, nous aurions les résultats d'appariement de la configuration : source1 -> référence / source2 -> cible et source1 -> cible /source2 -> référence. Généralement, le résultat sera l'intersection entre les ensembles issus des deux configurations.

3.3.4 Classification selon les étapes d'appariement

Les méthodes d'appariments sont composées de plusieurs étapes, et cela change d'une publication à une autre. Néanmoins, nous pouvons les réduire et les classifier en deux grandes stratégies :

- **Les méthodes en 2 phases** : Sélectionner/Filtrer/Réduire les candidats possibles puis décider sur la base de ces candidats lesquelles se correspondent.
- **Les méthodes en 1 phase** : Prendre la décision sur la base de tous les objets disponibles.

Les méthodes en 2 phases : sont les plus populaires [51, 36, 3, 50, 63, 10]. Elles aboutissent à l'appariement en deux phases : sélection des candidats et décision sur ces candidats. Il se peut que ces phases se répètent plusieurs fois : plusieurs sélections puis décision ou altérer entre plusieurs sélections et décisions. Comme dans [63], où ils utilisent un buffer pour choisir les candidats. Puis, ceux qui sont retenus devront satisfaire des critères géométriques. Ensuite, en cas de conflit (plusieurs choix) ils appliquent d'autres mesures topologiques.

- **Sélection des candidats** : consiste à réduire l'espace de recherche pour un objet (point, ligne ou polygone) pour lequel on cherche un(des) correspondant(s). Par conséquent, réduire le temps de calcul. Nous pouvons regrouper les outils pour sélectionner ces candidats en 3 catégories :

  - **Buffer** : est basé sur l'hypothèse que la notion de distance est présente entre les objets à apparier, Cette méthode consiste à construire un buffer autour de chaque objet de l'ensemble de référence pour lequel on cherche un correspondant. Ensuite, les objets de l'ensemble de comparaison qui se situent à l'intérieur de ce buffer sont considérés comme des candidats possibles. Par exemple, sur la figure 3.12, le buffer est construit autour de l'arête rouge numéro 99 de la source de référence. Et puisque les arêtes 1571, 1572 et 1573 de la source de comparaison sont à l'intérieur de ce buffer. Ils sont considérés comme des candidats possibles pour l'appariement.

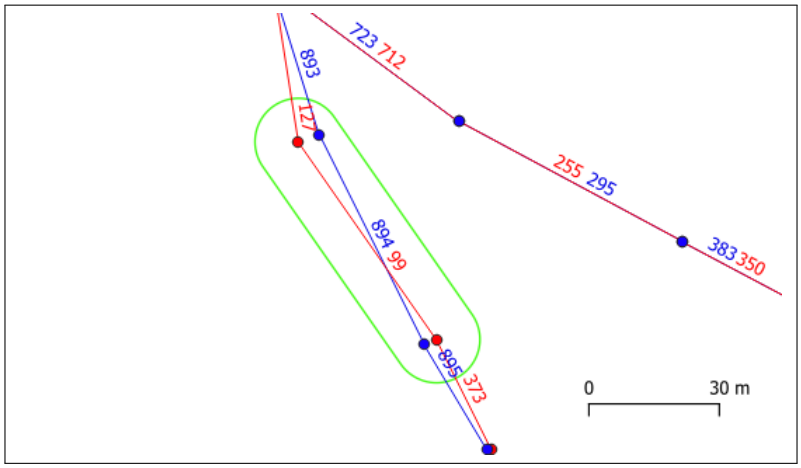


FIGURE 3.12 – Buffet de sélection

  - **Filtres** : consistent à définir des contraintes sur des mesures de similarité pour réduire le nombre des candidats possibles. Par exemple, dans [36] [Samal et al 2004] ils construisent une matrice de similarité à partir de plusieurs mesures : distances, formes et attributs. Puis, dans les étapes suivantes ils vont se servir que des objets qui ont une valeur maximale sur leurs colonnes et lignes (figure 3.13).

  - **Buffer + filtre** : consistent à appliquer un buffer sur les objets, puis utiliser les filtres pour éliminer les objets qui sont à l'intérieur de ce buffer et qui ne respectent pas des contraintes des mesures de similarités. Par exemple, dans le figure 3.14 un buffer est construit autour des objets. Ensuite, des contraintes sur

	Source 1					
Source 2	12	34	35	45	56	66
123	0.00	0.12	**0.76**	0.10	0.00	0.00
145	0.23	0.45	0.10	**0.88**	0.00	0.11
234	0.00	0.00	0.00	0.00	0.40	0.40
255	0.23	0.10	0.30	0.48	0.00	0.54
267	**0.65**	0.00	0.00	0.20	0.10	0.00
315	0.10	0.04	0.15	0.05	0.00	0.05

FIGURE 3.13 – Matrice de similarités [36]

des mesures de similarités ont été appliquer comme :

* L'angle entre l'objet de référence et le candidat ne doit pas dépasser un seuil fixé.
* La différence de la longueur de l'objet de référence et candidats ne doit pas dépasser un seuil fixé.

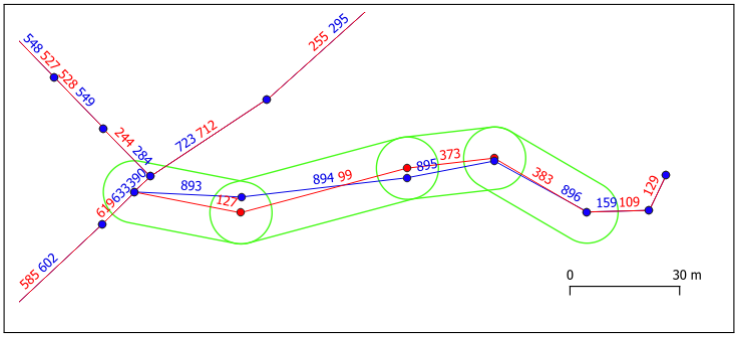


FIGURE 3.14 – Buffer de sélection

Le problème majeur de ces méthodes de sélection, est la taille ou le seuil à choisir pour le buffer ou les filtres. Si par exemple, on prend un buffer grand il va inclure beaucoup de fausses positives et sinon, nous allons trouver peu de correspondances entre les objets. Quelques publications ont essayé de surpasser ce problème en effectuant des études statistiques sur les données avant de fixer ces valeurs comme dans [51].

- **Décision** : l'étape de sélection produit comme sortie un ensemble restreint de candidats qui peuvent s'apparier entre eux, dont plusieurs qui sont en conflit. C'est-à-dire, des objets qui matchent avec des objets différents (Attention, ce qui rend la tâche difficile dans le cas des cardinalités 1 :n est de faire la distinction entre un conflit et un objet qui, en réalité, s'apparie avec plusieurs objets). Pour décider de ces candidats nous avons pu extraire deux approches différentes :

– Soit sur la base des mesures utilisées dans la sélection des candidats soient en définissants de nouveaux critères. Les conflits sont résolus en choisissant celui qui a le meilleur score en termes de ces mesures. Le score est calculé en faisant la somme du produit des mesures fois leur poids. Par exemple, dans le cas de deux sources de données, le score de similarité entre les objets a et b est :

$$S(a,b) = \sum_{1}^{m} \text{poids}(i) * \text{Mesure}(i,a,b) \tag{3.2}$$

Avec m le nombre de mesures de similarités utilisées. Cette approche est utilisée dans [49, 52].

– Soit en utilisant une fonction d'optimisation pour laquelle on donne aussi, soit les valeurs des critères issus de la sélection ou de nouvelles mesures. Dans [36], ils ont transformé ce problème à celui de la recherche des cliques maximales dans un graphe. La figure 3.15 illustre la modélisation de ce problème, tel que les noeuds ce sont les candidats, les arêtes ce sont les scores de similarités entre eux. L'objectif est de trouver dans le graphe le groupe qui maximise ces scores.

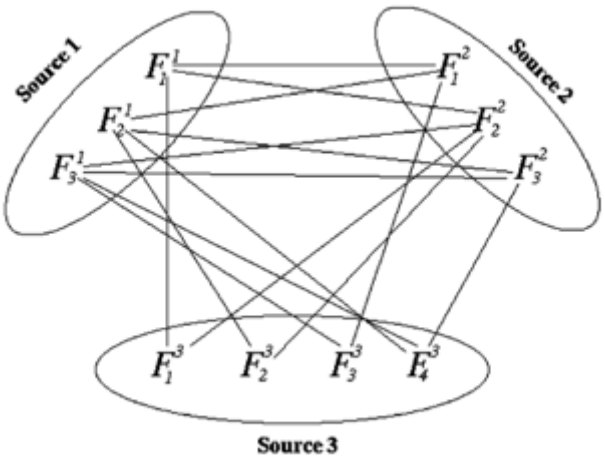


FIGURE 3.15 – Graphe de similarité entre trois sources [36]

Dans [51], ils ont modélisé ce problème sous forme de problème de communication : émetteur, canal et récepteur. L'émetteur c'est un objet de la source1, le récepteur est de la source2. L'objectif est de trouver le canal qui garde le maximum d'informations (similarités) entre l'émetteur et le récepteur.

Les méthodes en 1 phases : Cette stratégie apparie les objets directement sans la sélection des candidats. C'est-à-dire, ils considèrent tous les objets à la fois. Dans [3] ils calculent

pour tous les objets la probabilité qu'ils s'apparient en se basant sur la distance. Les couples retenus sont ceux qui avaient la probabilité la plus élevée. Dans [23], après avoir choisi les mesures de similarité. Ils modélisent l'appariement d'objet en problème d'affection, dont l'objectif est de minimiser la dissimilarité : $\mathrm{Min} \sum_{i=1}^{n} \sum_{j=1}^{m} C_{i,j} x_{i,j}$. Tel que, i l'index de l'objet dans la source1, j l'index de l'objet de la source2, n le nombre d'objets dans la source1, m le nombre d'objets dans la source2 et $C_{i,j}$ est la dissimilarité entre deux objets (on peut la calculer à partir des mesures de similarités) et $x_{i,j}$ valeur booléenne égale à 1 si les objets se correspondent et 0 sinon. Dans [24], ils cherchent à trouver tous les Matchs à la fois en maximisant une fonction objective : problème d'optimisation.

3.3.5 Classification Statistique / non-statistique

Il existe plusieurs travaux qui utilisent les statistiques dans leur approche à plusieurs niveaux. Par exemple, dans [51], comme indiqué précédemment ils appliquent un buffer et des filtres pour sélectionner les candidats possibles. Les valeurs de seuils de ces filtres sont obtenues en effectuant sur un échantillon de données une analyse de la fréquence d'un ensemble de mesures de similarités. Dans [3, 49, 52], ils utilisent les probabilités pour quantifier les similitudes entre les objets.

3.3.6 Classification Itérative / non-itérative

Les méthodes itératives apparient les objets sur plusieurs itérations. Dans [36], les auteurs utilisent un buffer qui augmente de taille au fur est à mesure que l'appariement est effectué. Dans [50], après avoir matcher les lignes qui ont des scores très élevés en termes de mesures de similarités, leurs adjacents seront appariés d'une manière itérative jusqu'au ce que les objets soient épuisés.

Chapitre 4

Contribution à la fusion des données imparfaites

4.1 Introduction

Dans l'objectif de maîtriser la gestion des imperfections lors de la fusion des données multi-sources, nous proposons dans ce chapitre une catégorisation de ces imperfections qui permettrait de les traiter de manière générique. Nous présentons ensuite une approche de fusion des données imparfaites pour l'appariement de données géographiques basée sur la théorie des fonctions de croyance [29], que nous allons implémenter et adapter à notre contexte.

4.2 Catégorisations des formes d'imperfections liées aux réseaux d'assainissement

Pour une meilleur gestion des imperfections liées aux réseaux d'assainissement, nous proposons une catégorisation des ces dernières permettant de les traiter de manière générique. Rappelons que dans les bases de données disponibles, les nœuds qui représentent réellement des ouvrages, des appareillages et des réparations (regards, avaloir etc.), et les arêtes qui illustrent les canalisations (collecte, transport, etc.) sont stockés de manière séparée dans des shapefiles. Ainsi, notre catégorisation sépare les formes d'imperfections liées aux nœuds d'une part, et aux arêtes d'autre part.

4.2.1 Formes d'imperfections liées aux nœuds

Nous considérons dans la suite de ce chapitre que les nœuds en rouge dans les figures représentent les éléments de la base de données de référence (dans notre cas la source officielle des données des réseaux d'assainissement de la ville de Montpellier), et que les nœuds en bleu représentent les éléments de la base de données de comparaison (dans notre cas, les nœuds qui sont détectés à partir des images haute résolution).

4.2.1.1 Décalage d'un nœud avec un seul candidat

Cette catégorie se caractérise par le fait qu'il n'y a qu'un seul nœud candidat de la base de données de comparaison à l'intérieur du buffer qui se trouve autour du nœud de référence, mais avec un décalage entre les deux nœuds, qui représentent en réalité le même objet réel, et

doivent donc être fusionnés. La forme d'imperfection présente dans cette catégorie d'imperfection est l'imprécision. Un exemple est donné dans la figure 4.1 où un seul candidat (270) est sélectionné pour le nœud de référence 648, mais avec un décalage entre l'objet de référence et l'objet de comparaison. Le même cas se présente pour les noeuds 762 et 594. les nœuds candidats doivent dans ce cas être fusionnés avec les noeuds de référence.

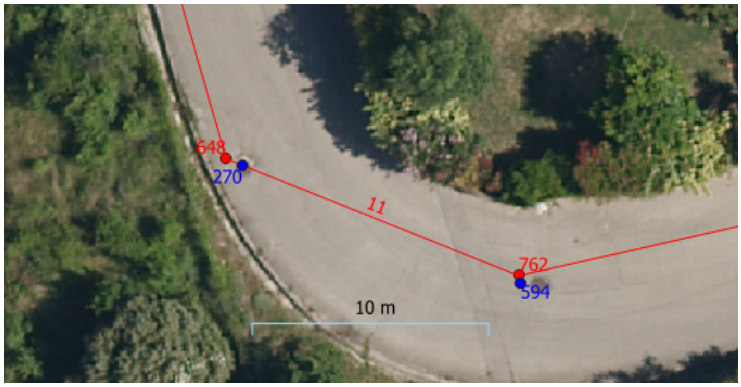


FIGURE 4.1 – Exemple de décalage d'un nœud avec un seul candidat

4.2.1.2 Décalage d'un nœud avec plusieurs candidats

Dans cette catégorie, plusieurs candidats sont sélectionnés pour un même nœud de référence mais avec un certain décalage dans leurs positions. En réalité, le nœud de référence doit correspondre à l'un des nœuds candidats ou ne correspondre à aucun d'entre eux. Après appariement, certains nœuds candidats resteront sans correspondants et seront considérés comme incomplets dans la base de données de référence. Ainsi, cette catégorie concerne les deux aspects d'imperfection, à savoir, l'imprécision et l'incomplètude des données.

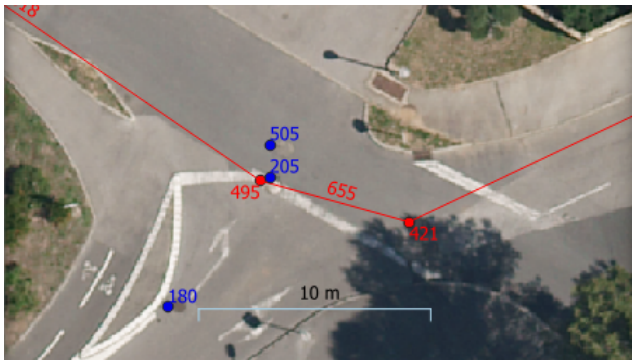


FIGURE 4.2 – Exemple de décalage d'un nœud avec plusieurs candidats

L'exemple de la figure 4.2 montre le nœud de référence numéro 495 qui possède deux candidats dans la base de données de comparaison 205 et 505. En fait, le nœud 495 devrait être

fusionné avec le nœud de comparaison numéro 205 et le nœud de comparaison numéro 505 devrait être considéré comme incomplet dans la base de données de référence.

4.2.2 Formes d'imperfections liées aux arêtes

Nous considérons dans ce qui suit que les arêtes en rouge sont les éléments de la base de données de référence (dans notre cas la source officielle des données des réseaux d'assainissement de la ville de Montpellier), et les arêtes en bleu sont les éléments de la base de données de comparaison (dans notre cas la source des données à partir de l'analyse d'images haute résolution).

4.2.2.1 Décalage uniforme de type 1 :1

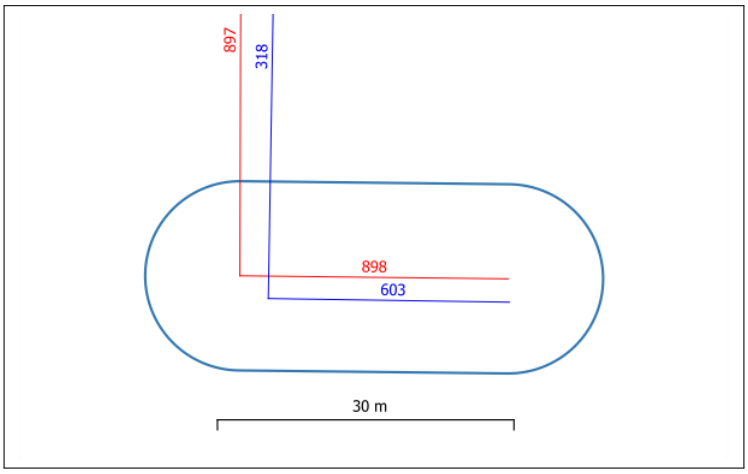


FIGURE 4.3 – Exemple de décalage uniforme entre les lignes avec une cardinalité de 1 :1

Cette catégorie considère qu'une arête de référence peut avoir plusieurs candidats mais avec une et une seule arête candidate qui représente le correspondant réel. Nous considérons ici le décalage uniforme entre l'arête de référence et l'arête cible, c'est à dire que les arêtes sont parallèles (angle entre ces deux arêtes égale à 0). Les deux arêtes, de référence et cible, doivent être fusionnées avec une cardinalité de type 1 :1. Le type d'imperfection présent dans cette catégorie est l'imprécision des données. En prenant l'exemple de la figure 4.3, l'arête de référence numéro 898 a deux candidats 603 et 318 avec un décalage uniforme entre l'arête de référence et le candidat numéro 603. Ainsi, l'arête de référence doit être appareillé avec le seul candidat en décalage uniforme qui est l'arête de comparaison numéro 603.

4.2.2.2 Décalage non uniforme de type 1 :1

Dans cette catégorie, il est possible d'avoir plusieurs candidats pour l'arête de référence, mais avec un décalage non uniforme entre l'arête de référence et ces candidats. C'est-à-dire que l'angle entre eux est différent de 0. Un seul candidat représente le correspondant réel de l'arête de référence (cardinalité 1 :1). La forme d'imperfection considérée ici est l'imprécision des données. Dans l'exemple de la figure 4.4, l'arête de référence numéro 99 possède trois candidats 893, 894 et 895. Le candidat le plus proche est 894 avec laquelle l'arête de référence doit être appareillée.

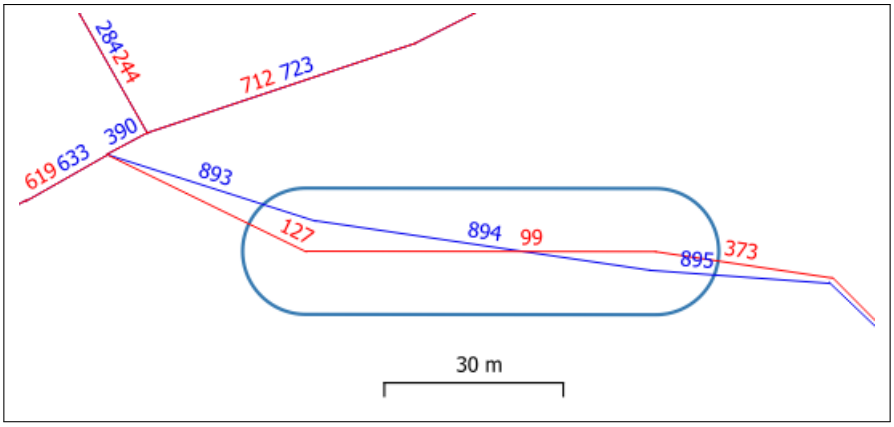


FIGURE 4.4 – Exemple de décalage non uniforme entre les lignes avec une cardinalité de 1 :1

4.2.2.3 Décalage uniforme de type 1 :n

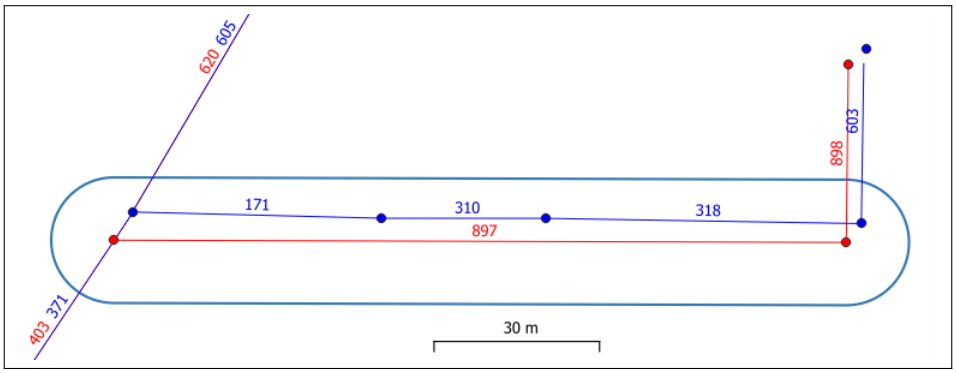


FIGURE 4.5 – Exemple de décalage uniforme entre les lignes avec une cardinalité de 1 :n

Dans cette catégorie, plusieurs candidats existent pour l'arête de référence. Cette dernière correspond à une collection d'arêtes candidates avec lesquelles le décalage est uniforme, c'est à dire que l'angle entre ces candidats et l'arête de référence est égal à 0. La cardinalité à considérer dans cette catégorie est 1 :n et le type d'imperfection est l'imprécision des données.

Dans l'exemple de la figure 4.5, l'arête de référence numéro 897 a plusieurs candidats, mais entre l'arête de référence et les candidats numéros 171, 310 et 318 il y a un décalage uniforme et la somme de la longueur de ces trois candidats est presque égale à la longueur de l'arête de référence numéro 897. L'arête de référence doit donc être appareillée avec les trois candidats de la base de données de comparaison numéros 171, 310 et 318.

4.2.2.4 Décalage non uniforme de type 1 :n

Dans cette catégorie plusieurs candidats existent pour l'arête de référence. Cette dernière correspond à une collection d'arêtes candidates avec lesquelles le décalage est non uniforme, c'est à dire que l'angle entre eux est différent de 0. La cardinalité considérée dans cette catégorie est 1 :n, et le type d'imperfection est l'imprécision des données. Dans l'exemple de la figure 4.6, l'arête de référence numéro 1100 a plusieurs candidats, mais entre l'arête de référence et les candidats numéros 1201, 1202 et 1203 il y a un décalage non uniforme. Ajoutons à cela que la somme de la longueur de ces trois candidats est presque égale à la longueur de l'arête de référence numéro 1100, l'arête de référence doit donc être fusionnée avec les trois candidats de la base de données de comparaison numéros 1201, 1202 et 1203.

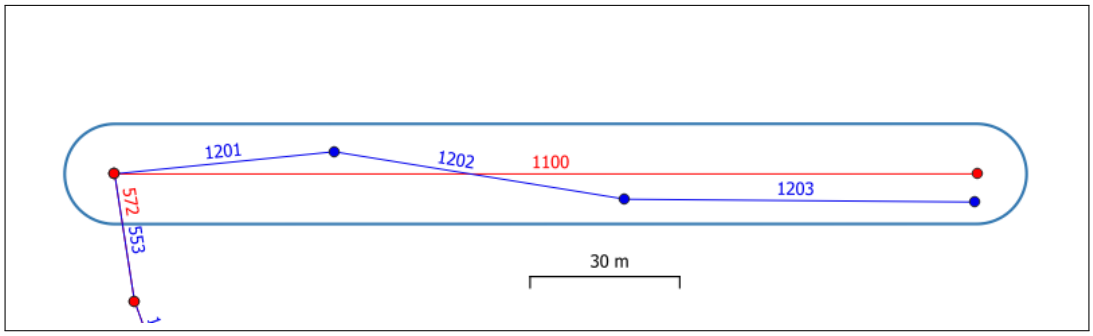


FIGURE 4.6 – Exemple de décalage non uniforme entre les lignes avec une cardinalité de 1 :n

4.3 Le processus d'appariement de données géographiques selon [29]

4.3.1 Choix de l'approche

L'approche proposée par [29] concerne l'appariement des réseaux routiers(données linéaires) et les reliefs (données ponctuelles). Dans ce travail, l'initialisation des masses de croyances est effectuée sur la base du modèle proposé par Appriou [2], dans lequel chaque source se spé-

cialise dans une hypothèse du cadre de discernement. C'est-à-dire que la source d'information analyse l'hypothèse donnée et se prononce en faveur de celle-ci, contre celle-ci ou ne se prononce pas si elle n'a pas assez de connaissances sur cette hypothèse. Nous jugeons que cette approche est pertinente pour notre cas puisque sa modélisation des masses permet de quantifier :

- la connaissance complète.
- la connaissance incomplète.
- l'ignorance.

4.3.2 Principe général

Dans cette section, nous présentons le processus d'appariement selon [29]. Il se compose de cinq étapes principales, explicitées dans la figure 4.7.

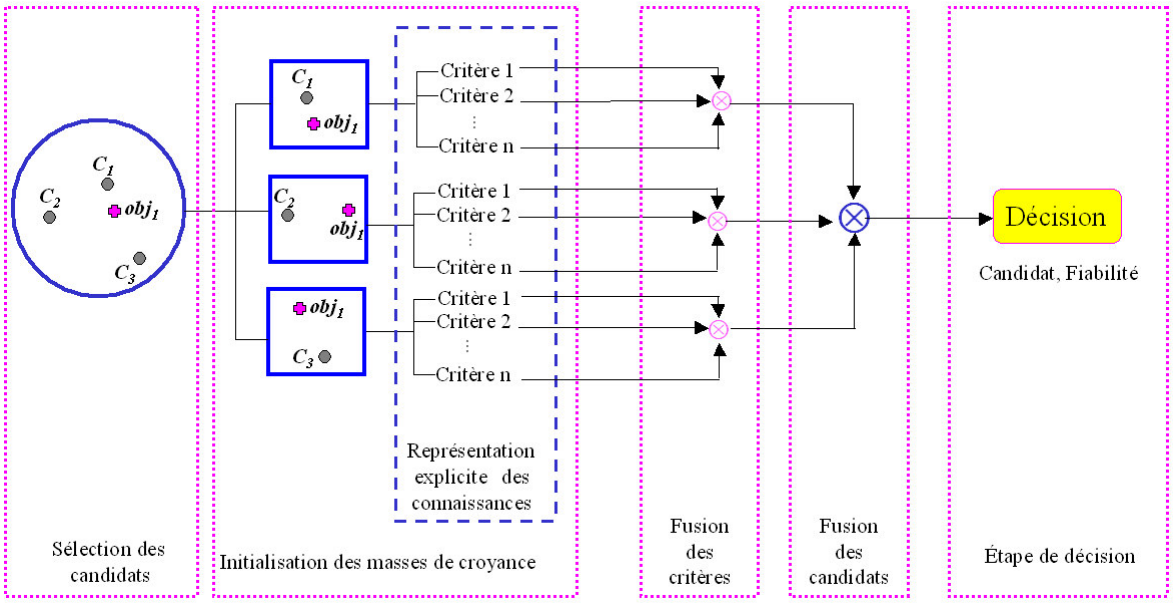


FIGURE 4.7 – Processus d'appariement de données géographiques détaillé [29]

Le processus d'appariement utilise deux bases de données BD1 et BD2, chaque objet obj de BD1 est examiné, tour à tour, avec les objets de BD2 pour trouver son homologue. La première étape correspond à la sélection des candidats, qui consiste à rechercher, pour chaque objet appartenant à BD1, des objets homologues potentiels dans BD2, appelés candidats d'appariement et notés $C_{i,(i=1,2,...,N)}$. Ensuite, chacun des candidats d'appariement est analysé pour déterminer le lien de correspondance.

L'initialisation des masses de croyances (deuxième étape du processus) consiste, pour chaque critère d'appariement, à se prononcer sur chacun des candidats en affectant une croyance aux hypothèses définies pour chaque candidat. Les critères d'appariement peuvent être la distance euclidienne ,le toponyme ou l'orientation, Ensuite, et après avoir fusionné les critères d'appariement pour chaque candidat (troisième étape du processus), la fusion des candidats est effectuée pour avoir une vue globale des croyances attribuées à tous les candidats (quatrième étape du processus). Finalement, l'étape de décision consiste à choisir le meilleur candidat.

4.3.3 Sélection des candidats

Le processus commence par la sélection des candidats pour l'appariement (voir figure 4.8).

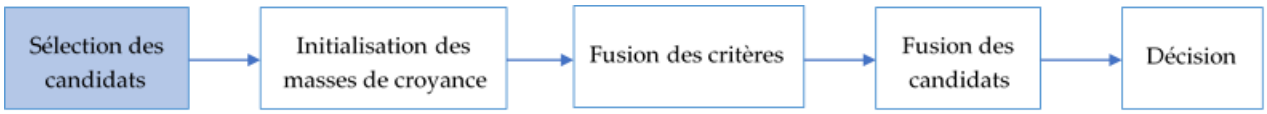


FIGURE 4.8 – Etape de sélection des candidats à l'appariement

La sélection des candidats permet de définir le cadre de discernement associé à un objet $obj1$ de la base de données de référence. Pour ce faire, les objets $(obj2i), i = 1 \ldots N$ les plus proches dans la base de données BD2 selon un critère de distance géométrique sont considérés. Par exemple, dans la figure 4.9, les candidats d'appariement de l'objet de référence numéro 525 selon la distance euclidienne avec un seuil de 3 mètres sont 166, 228, 386, 541 et 598.

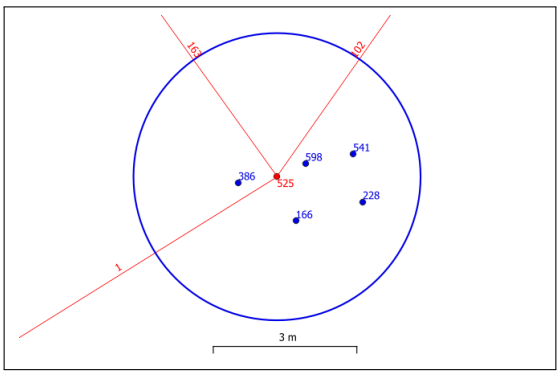


FIGURE 4.9 – Candidats à l'appariement pour l'objet 525 sélectionnés selon un critère de distance géométrique

En général, la détermination du seuil de la distance dépend de la connaissance de l'erreur moyenne sur la position et de la nature des objets. Par conséquent, un seuil de sélection

élevé permettra de prendre en compte et gérer l'imprécision sur la localisation des objets géographiques. Par contre, un seuil très élevé augmentera la complexité algorithmique. En effet, plus le seuil est élevé, plus il y a de candidats et plus il y a de combinaisons possibles dans 2^{Θ}.

L'identification de ces candidats d'appariement permet ainsi de définir le cadre de discernement qui considère que l'appariement de l'objet $obj1$ à chacun des candidats choisis constitue une hypothèse, et donc une solution possible. Dorénavant, nous étiquetons chaque objet candidat $(obj2i), i = 1 \ldots N$ avec $Ci, i = 1 \ldots N$.

Ainsi, le cadre de discernement est défini comme un ensemble d'hypothèses $C_i, (i = 1, 2 \ldots N)$, chacune exprime le fait que "le candidat Ci est l'homologue de l'objet $obj1$". Cependant, des objets peuvent ne pas avoir de correspondants dans la seconde base de données et par conséquent ne peuvent pas être appariés. Une nouvelle hypothèse a été ajoutée (appelée NA) qui signifie que "l'objet $obj1$ n'est pas apparié". L'ajout de l'hypothèse NA permet de rendre le cadre de discernement exhaustif, autrement dit, la solution est toujours trouvée parmi les hypothèses définies. Grâce à l'hypothèse NA, le conflit après la fusion des critères devient alors moins important parce que la réponse se trouve toujours dans le cadre de discernement.

Pour un objet de la base de données de référence avec N candidats à l'appariement, le cadre de discernement est défini par :

$$\Theta = \{C_1, C_2, \ldots, C_i, \ldots, C_N, NA\} \tag{4.1}$$

A partir de ce cadre de discernement, le référentiel de définition 2^{Θ} est determiné. Il comporte alors toutes les combinaisons possibles des hypothèses définies dans le cadre de discernement. En effet, pour un cadre de discernement contenant N hypothèses provenant de la sélection de N candidats, auxquelles on ajoute l'hypothèse NA, on a 2^N combinaisons.

$$2^{\Theta} = \{C_1, \ldots, C_N, \{C_1, C_2\}, \{C_1, C_3\}, \ldots, \{C_1, \ldots, C_N, \mathrm{NA}\}, \mathrm{NA}\} \tag{4.2}$$

4.3.4 Initialisation des masses de croyance

La deuxième étape consiste à initialiser les masses de croyance pour chaque candidat, indépendamment des autres candidats, et pour chaque source d'information. Cette étape est une tâche cruciale dans le processus d'appariement, qui agit sur la qualité des résultats atten-

dus.

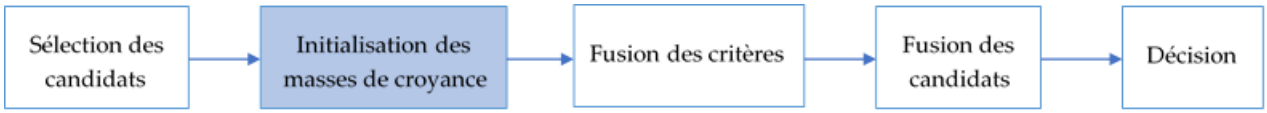


FIGURE 4.10 – Deuxième étape du processus : initialisation des masses de croyance

Dans l'approche proposée par [29], l'initialisation des masses de croyances est basée sur des fonctions floues définies à partir des connaissances. L'implémentation est réalisée en utilisant l'un des modèles proposés par [2] qui considère le sous-ensemble S_i de 2^{Θ}, défini comme suit :

$$S_i = \{C_i, \neg C_i, \Theta\} \tag{4.3}$$

Avec :

- C_i : Représente l'hypothèse selon laquelle l'objet de référence analysé est apparié avec le candidat C_i, ce qui signifie une connaissance complète.
- $\neg C_i = \{C_1, \ldots, C_{i-1}, C_{i+1}, \ldots, C_N\}$: Représente l'hypothèse selon laquelle l'objet de référence analysé est apparié avec un candidat autre que C_i ou n'est pas apparié du tout, ce qui signifie une connaissance incomplète.
- $\Theta = \{C_1, C_2, \ldots, C_i, \ldots, C_N, \mathrm{NA}\}$: représente l'hypothèse selon laquelle on ne peut pas se prononcer sur ce candidat, ce qui signifie l'ignorance.

Il est à noter que la somme des poids de croyance attribués aux trois hypothèses doit être égale à 1 : $m(C_i) + m(\neg C_i) + m(\Theta) = 1$.

4.3.5 Fusion des critères

Une fois les masses de croyance initialisées pour chacun des candidats C_i, les ensembles de masses de chaque critère d'appariement sont fusionnés avec l'opérateur conjonctif. Autrement dit, l'opérateur de Dempster est appliqué sans normalisation des masses de croyance (voir Figure 4.11). Finalement, pour chacun des candidats C_i, un ensemble de masses composé des hypothèses $C_i, \neg C_i, \Theta$ est obtenu synthétisant les avis de tous les critères qui ont été prononcés individuellement.

Par exemple, dans la figure 4.12, deux critères d'appariement (Critère 1 et Critère 2) sont considérés. Ils se prononcent sur chaque candidat C_i. Notons que $\{m_1(C_i), m_1(\neg C_i), m_1(\Theta)\}$

FIGURE 4.11 – Troisième étape du processus : fusion des critères

est l'ensemble des masses défini par le Critère 1 et $\{m_2(C_i), m_2(\neg C_i), m_2(\Theta)\}$ est l'ensemble des masses défini par le Critère 2. Afin de prendre en compte les avis des deux critères, les deux ensembles de masses sont fusionnés à l'aide de l'opérateur conjonctif. Le jeu de masses obtenu après la fusion des critères est le suivant $\{m_{12}(C_i), m_{12}(\neg C_i), m_{12}(\Theta), m_{12}(\emptyset)\}$. la masse de croyances $m_{12}(\emptyset)$ représente le conflit généré par le désaccord entre les connaissances de chaque critère.

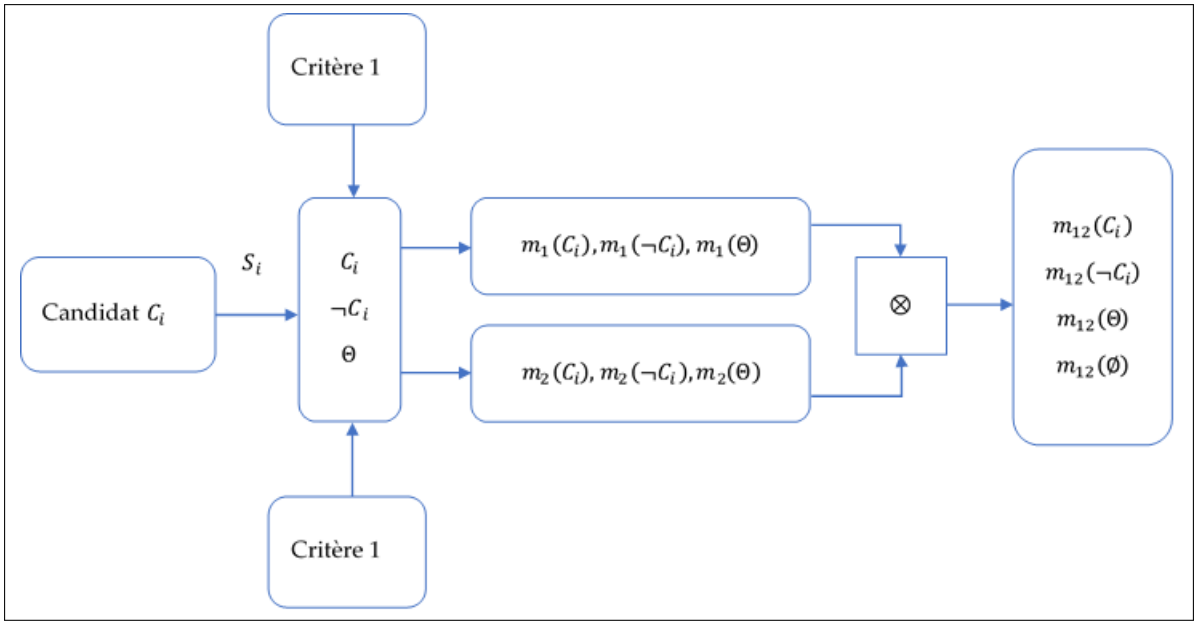


FIGURE 4.12 – Fusion des critères pour le candidat C_i

4.3.6 Fusion des candidats

La quatrième étape du processus d'appariement est la fusion des candidats (voir figure 4.13).

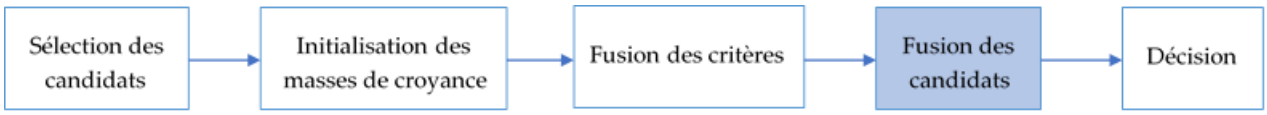


FIGURE 4.13 – Quatrième étape du processus : fusion des candidats

Au cours de cette étape (voir figure 4.14), une méthode de fusion proposée par [33] dans le contexte d'une application de localisation d'un véhicule sur une carte a été adoptée. La fusion des candidats est réalisée en fusionnant avec l'opérateur conjonctif, l'ensemble des masses du candidat C_1 avec l'ensemble des masses du candidat C_2. Ensuite, le résultat avec l'ensemble des masses du candidat C_3, et ainsi de suite jusqu'au candidat N. Rappellons que, l'opérateur conjonctif étant associatif, peu importe l'ordre de fusion des candidats. La masse de croyance obtenue après avoir fusionné les N candidats est notée $m_{1\ldots N}$.

La masse de croyances allouée à l'hypothèse NA provient de la fusion de toutes les masses de croyances allouées aux hypothèses $\neg C_i$, pour $i = 1 \ldots N$. Elle est définie de la façon suivante :

$$m_{1\ldots N}(\mathrm{NA}) = m_{1\ldots N}(\neg C_1) * m_{1\ldots N}(\neg C_2) * \cdots * m_{1\ldots N}(\neg C_i) * \cdots * m_{1\ldots N}(\neg C_N) \tag{4.4}$$

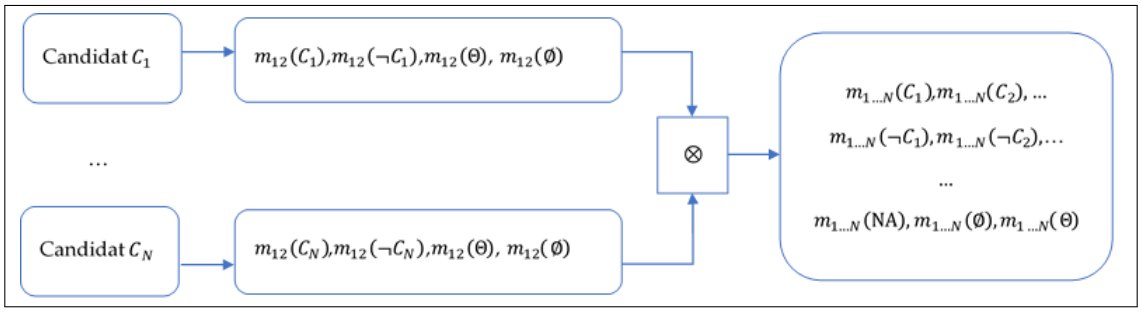


FIGURE 4.14 – Combinaison des masses de croyance pour chacun des candidats

4.3.7 Décision

La cinquième étape du processus concerne la prise de décision (voir figure 4.15).

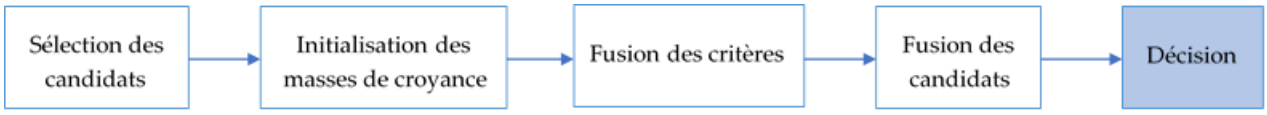


FIGURE 4.15 – Cinquième étape du processus : décision

Cette décision est prise après la phase de fusion des candidats et après avoir normalisé les masses résultantes à l'aide de l'opérateur de Dempster. Le conflit est égal à 0, car il est utilisé pour la normalisation des masses de croyances.

Rappelons que le premier objectif était d'apparier un objet à un seul candidat. Par conséquent, cette approche permettrait d'apparier d'une part des objets possédant un seul homologue dans l'autre base de données, mais également des objets possédant plusieurs homologues. Cette dernière fonction peut être réalisée en regroupant les correspondances obtenues.

Ainsi, l'utilisation de la probabilité pignistique dans l'étape de décision paraît approprié puisque la probabilité pignistique favorise les hypothèses simples.

Pour prendre une décision, la probabilité pignistique maximale est utilisée, c'est-à-dire l'hypothèse simple qui a la masse de croyance la plus élevée. Par exemple, la figure 4.16 indique qu'à partir de l'ensemble des masses obtenues après la fusion des candidats, les probabilités pignistiques de toutes les hypothèses simples sont calculées, pour ensuite sélectionner celle qui a la probabilité pignistique maximale. Dans notre exemple, il s'agit de l'hypothèse C_i signifiant que le candidat homologue de l'objet de référence est le candidat C_i.

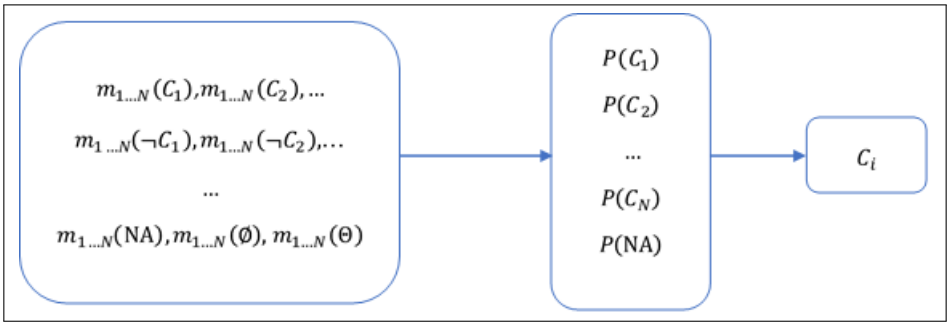


FIGURE 4.16 – Décision selon le maximum de probabilité pignistique

4.4 Adaptation de l'approche à notre contexte

4.4.1 Les bases de données utilisées

Les données des réseaux d'assainissement sont composées de nœuds et d'arêtes, leur type est présenté dans l'annexe A.3. Pour le stockage de ces données on utilise les shapefiles voir l'annexe A.4, et pour les visualiser nous utilisons le logiciel QGIS [1] (voir annexe A.2). Enfin, pour manipuler ces données en python en utilisant la bibliothèque GDAL [2] voir l'annexe A.5.

Les bases de données disponibles des réseaux d'assainissement sont :

1. *Quantum Geographic Information System*
2. *Geospatial Data Abstraction Library*

- La source 3M : Il s'agit de la source officielle de données géographiques pour la ville de Montpellier dans laquelle on trouve les données des réseaux d'assainissement (nœuds et arêtes).
- Nœuds détectés à partir des images à très haute résolution : Dans [8], les auteurs proposant une méthode de détection et de localisation automatique des nœuds dans des images aériennes et de télédétection à très haute résolution en utilisant un réseau neuronal convolutif (CNN)(voir figure 4.17).

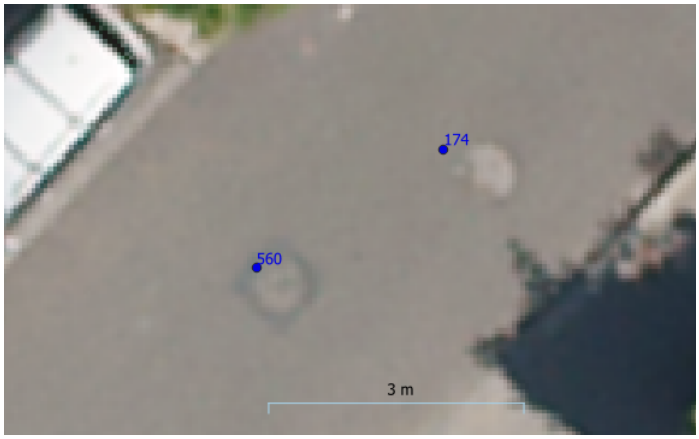


FIGURE 4.17 – Exemple de noeuds détectés à partir d'images à haute résolution

Notre objectif est d'étudier les imperfections présentes dans la fusion de deux bases de données de nœuds en considérant la source 3M comme la base de données de référence et la source des nœuds détectés à partir des images à haute résolution comme la base de données de comparaison, et notre but est de compléter et d'enrichir la base de données de référence. Par contre, pour les arêtes, nous ne disposons que d'une seule base de données. Par conséquent, nous avons créé les imperfections que l'on peut rencontrer en fusionnant deux sources d'arêtes et en les étudiant.

4.4.2 Initialisation des masses de croyance

L'étape d'initialisation des masses de croyance permet de définir les fonctions de masses associés aux différents critères considérés. Dans cette section, nous présentons les critères d'appariement que nous avons adoptés dans le contexte de l'appariement des objets dans un réseau d'assainissement. Nos critères s'expriment sous forme de seuils ou de distances, qui ont besoin d'être représentés explicitement dans le contexte de la théorie des fonctions de croyance, c'est-à-dire l'initialisation d'ensembles de masses de croyance pour chaque critère. Les courbes pré-

sentées ci-dessous et les seuils choisis sont des exemples typiques. L'ensemble du processus de modélisation des critères de correspondance dépend des données utilisées.

4.4.2.1 La géométrie des objets géographiques

Tout objet géographique possède une géométrie qui peut être représentée généralement par l'une des trois primitives suivantes : point, ligne ou polygone. La géométrie pour les points désigne donc la localisation, la forme pour la ligne et le polygone, mais aussi les relations spatiales implicites entre les objets géographiques. La géométrie des objets géographiques est une connaissance qui est toujours présente dans les bases de données géographiques et donc toujours exploitable.

De nombreuses mesures permettent d'exploiter et de comparer la géométrie des objets géographiques. Par exemple, pour comparer des localisations, il est possible d'utiliser une mesure de distance comme la distance euclidienne ou la distance de Hausdorff. Lorsque l'on doit comparer des lignes, on peut analyser leur orientation.

4.4.2.2 Le critère de la distance euclidienne

Dans le cadre de la fusion des données des réseaux d'assainissement, le critère d'écart de position a démontré son utilité avec la distance euclidienne appliquée aux nœuds qui représentent des ouvrages, des appareillages et des réparations (regards, avaloir etc.). La représentation de la connaissance explicite pour le critère de la distance euclidienne est présentée dans la figure 4.18.

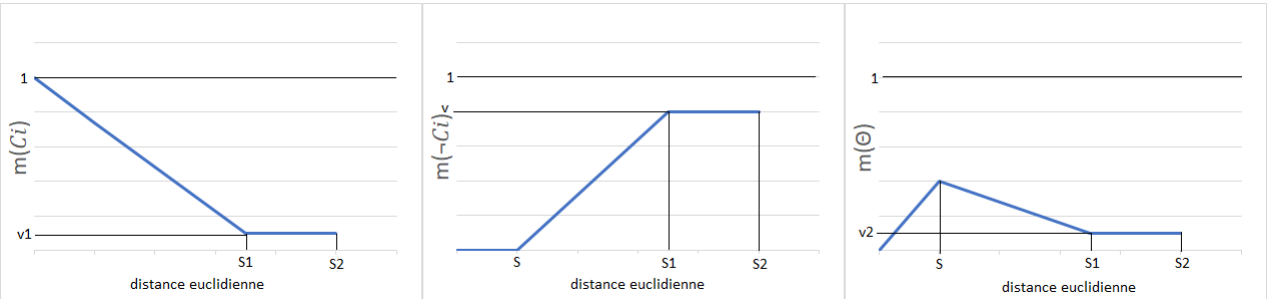


FIGURE 4.18 – Représentation des connaissances pour le critère de la distance euclidienne

La première courbe de la figure 4.18 représente la masse de croyance affectée à l'hypothèse C_i, c'est-à-dire "le candidat C_i est l'homologue de l'objet de référence". Cela signifie que plus un candidat C_i est proche de l'objet de référence, plus la probabilité que ce dernier

soit l'objet homologue est élevée. Elle signifie également qu'à partir d'un seuil $S1$ donné, nous considérons que cette hypothèse devient presque impossible. Pour éviter l'élimination définitive d'un candidat relativement éloigné, la masse de croyance affectée à cette hypothèse, $m(C_i)$ est égale à $v1$. Précisons qu'au-delà d'une distance égale à $S2$, c'est effectivement improbable puisque nous ne sommes plus dans le cadre du discernement. Cela signifie que nous cherchons des candidats qui ont une distance euclidienne avec l'objet de référence inférieur à $S2$.

La deuxième courbe représente la masse de croyance affectée à l'hypothèse $\neg C_i$, ce qui signifie que "le candidat C_i n'est pas l'homologue de l'objet de référence". Pour modéliser l'imprécision de la localisation des objets géographiques, nous introduisons un second seuil S. Lorsque la distance est inférieure au seuil S, nous considérons cette hypothèse très improbable, mais lorsque la distance est entre S et $S1$, l'hypothèse est progressivement crédible. Enfin, lorsque la distance est entre $S1$ et $S2$ l'hypothèse devient très crédible, et donc la masse attribuée à $\neg C_i$ est une valeur v forcément proche de 1.

Enfin, la dernière courbe représente la masse de croyance attribuée à l'ignorance, $m(\Theta)$. Pour le critère de la distance euclidienne, l'ignorance est importante quand la distance est au voisinage de S, c'est-à-dire lorsque le candidat ne se trouve ni assez loin pour conclure avec certitude que ce n'est pas lui, ni assez près pour conclure qu'il est le véritable homologue.

4.4.2.3 Le critère de l'angle

Le critère d'angle est illustré dans la figure 4.20. Ce critère est appliqué aux arêtes qui illustrent les canalisations (collecte, transport, etc.) des réseaux d'assainissement. Dans cet exemple, il consiste à calculer l'angle entre l'objet de référence et le candidat C_i.

Chaque arête se caractérise par deux points d'extrémité. Pour calculer l'angle entre l'arête de référence et l'arête candidate, nous utilisons le produit scalaire entre deux vecteurs :

$$\overrightarrow{AB} \cdot \overrightarrow{CD} = AB * BC * \cos(\theta) \tag{4.5}$$

Dans la figure 4.19, chaque arête est considérée comme un vecteur. La direction des arêtes n'est pas un élément important dans notre cas. Mais la formule du produit scalaire prend en compte les directions (les valeurs de $\cos(\Theta)$ sont comprises entre -1 et 1). Nous proposons de transformer les valeurs négatives en valeurs positives pour ne pas tenir compte des directions

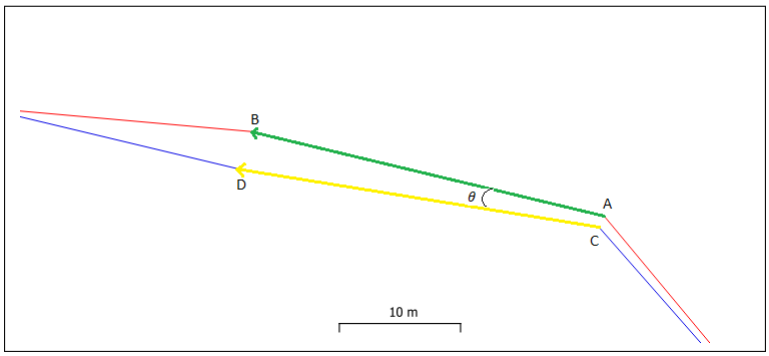


FIGURE 4.19 – Produit scalaire entre deux vecteurs

et donc, après cette étape, les valeurs de $\cos(\Theta)$ sont comprises entre 0 et 1. Cela signifie que les valeurs finales des angles appartiennent à l'intervalle $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

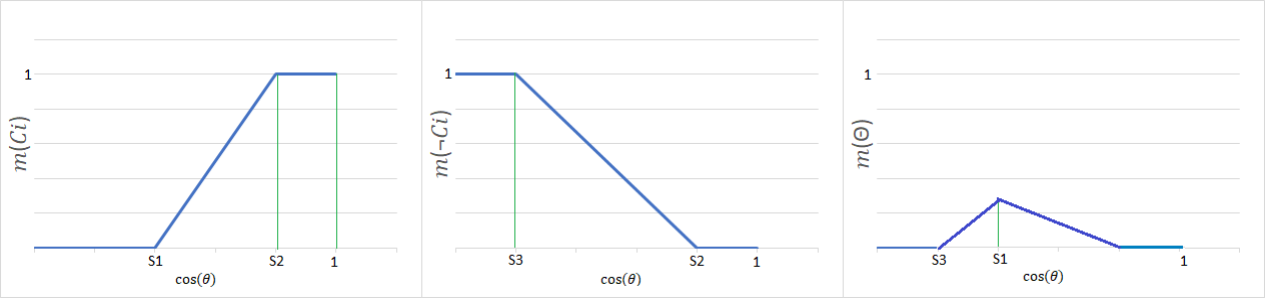


FIGURE 4.20 – Représentation des connaissances pour le critère de l'angle

La première courbe de la figure 4.20 représente la masse de croyance attribuée à l'hypothèse C_i, c'est-à-dire "le candidat C_i est l'homologue de l'objet de référence". En effet, si $cos(\Theta)$ est entre $S2$ et 1, nous supposons que les deux arêtes sont presque l'une sur l'autre, donc la masse assignée à C_i est égale à 1. Entre $S1$ et $S2$, plus l'angle entre l'objet de référence et l'objet de comparaison est petit, plus la probabilité que ce dernier soit l'objet homologue est élevée. Egalement, entre 0 et $S1$, nous considérons que l'angle entre les deux arêtes est très grand et donc cette hypothèse devient impossible.

La deuxième courbe représente la masse de croyance attribuée à l'hypothèse $\neg C_i$, qui signifie "le candidat C_i n'est pas l'homologue de l'objet de référence". Pour modéliser l'imprécision, nous considérons que si $cos(\Theta)$ est compris entre 0 et $S3$, l'hypothèse devient très crédible, et donc la masse attribuée à $\neg C_i$ est alors égale à 1, lorsque $cos(\Theta)$ est compris entre $S3$ et $S2$, l'hypothèse est de plus en plus improbable. Enfin, lorsque $cos(\Theta)$ est supérieur au seuil $S2$, nous considérons cette hypothèse comme très improbable.

Finalement, la dernière courbe représente la masse de croyance attribuée à l'ignorance, $m(\Theta)$. Pour le critère de l'angle, l'ignorance est importante lorsque l'angle est au voisinage de

$S1$, c'est-à-dire lorsque l'angle entre l'objet de référence et l'objet de comparaison n'est ni assez grand pour conclure avec certitude que ce n'est pas lui, ni assez petit pour conclure qu'il s'agit du véritable homologue.

4.4.2.4 Le critère de la distance de Hausdorff

Le critère de la distance de Hausdorff est illustré dans la figure 4.21. Ce critère est un outil topologique qui mesure la distance entre deux sous-ensembles d'un espace métrique sous-jacent. Cependant, les arêtes des réseaux d'assainissement sont caractérisées par deux points d'extrémité. Nous utilisons donc la distance de Hausdorff modifiée (MHD) entre deux ensembles de points S et T qui contiennent successivement les deux points d'extrémité de l'arête de référence et de l'arête de comparaison, cette distance est définie par :

$$\mathrm{M\,H\,D}(S,T) = \max\{g_d(S,T), g_d(T,S)\} \tag{4.6}$$

Où d une distance quelconque ; dans notre cas, nous utilisons la distance euclidienne. g_d est la distance de Hausdorff modifiée relative. Elle est définie par :

$$g(S,T) = \frac{1}{|S|}\sum_{p\in S}\min_{q\in T}\{d(p,q)\} \tag{4.7}$$

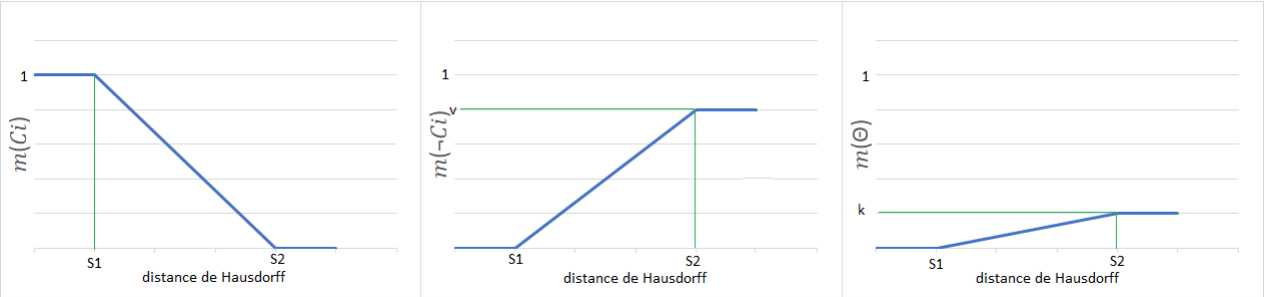


FIGURE 4.21 – Représentation des connaissances pour le critère de la distance de Hausdorff

La première courbe de la figure 4.21 représente la masse de croyance affectée à l'hypothèse C_i, c'est-à-dire "le candidat C_i est l'homologue de l'objet de référence". Cela signifie que plus un candidat C_i est proche de l'objet de référence, plus la probabilité que ce dernier soit l'objet homologue est élevée. Elle signifie également qu'à partir d'un seuil $S2$ donné, nous considérons que cette hypothèse devient impossible.

La deuxième courbe représente la masse de croyance affectée à l'hypothèse $\neg C_i$, ce qui

signifie "le candidat C_i n'est pas l'homologue de l'objet de référence". Lorsque la distance est inférieure au seuil $S1$, nous considérons cette hypothèse improbable, mais lorsque la distance est entre $S1$ et $S2$, l'hypothèse est progressivement crédible. Enfin, lorsque la distance est supérieure à $S2$ l'hypothèse devient très crédible, et donc la masse attribuée à $\neg C_i$ est une valeur v forcément proche de 1.

Enfin, la dernière courbe représente la masse de croyance attribuée à l'ignorance, $m(\Theta)$. Pour le critère de la distance de Hausdorff, l'ignorance prend un peu d'importance lorsque la distance de Hausdorff est au voisinage de $S2$.

4.4.2.5 Le critère du score

Rappelons que l'une des sources de données que nous considérons dans ce travail concerne la détection des plaques d'égouts à partir d'images à haute résolution en utilisant des techniques d'apprentissage automatique. En effet, dans [8], les plaques d'égout (noeuds du réseaux d'assainissement) sont automatiquement détectées sur des images à très haute résolution spatiale (5cm) grâce à un réseau de neurones convolutionnel. Le CNN AlexNet [22] a été adapté et entraîné sur une image de la commune de Prades-le-Lez (Hérault, France) contenant 605 plaques d'égout, puis validé sur les 101 plaques de la commune de Gigean (Hérault, France). En considérant les objets détectés comme plaque d'égout si le score de détection est supérieur à 0.9, les résultats ont un rappel de 50% et une précision de 68.5%. Nous proposons d'utiliser ce score de détection comme critère pour l'appariement des objets avec une autre source.

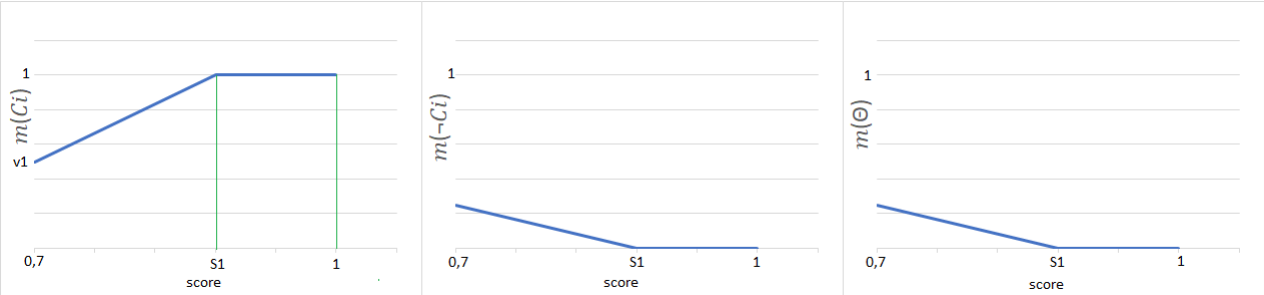


FIGURE 4.22 – Représentation des connaissances pour le critère du score

La première courbe de la figure 4.22 représente la masse de croyance affectée à l'hypothèse C_i, c'est-à-dire "le candidat C_i est l'homologue de l'objet de référence". Cela signifie que plus le score est grand, plus la probabilité que ce dernier soit l'objet homologue est élevée.

Cela signifie également qu'à partir d'un seuil $S1$ qui doit être très proche de 1, on considère que cette hypothèse devient très crédible.

La deuxième courbe représente la masse de croyance affectée à l'hypothèse $\neg C_i$, ce qui signifie "le candidat C_i n'est pas l'homologue de l'objet de référence". Et la dernière courbe représente la masse de croyance attribuée à l'ignorance, $m(\Theta)$. Nous affectons une masse considérable aux les hypothèses non $\neg C_i$ et Θ si le score est proche de 0,7.

4.4.2.6 Le critère de la longueur

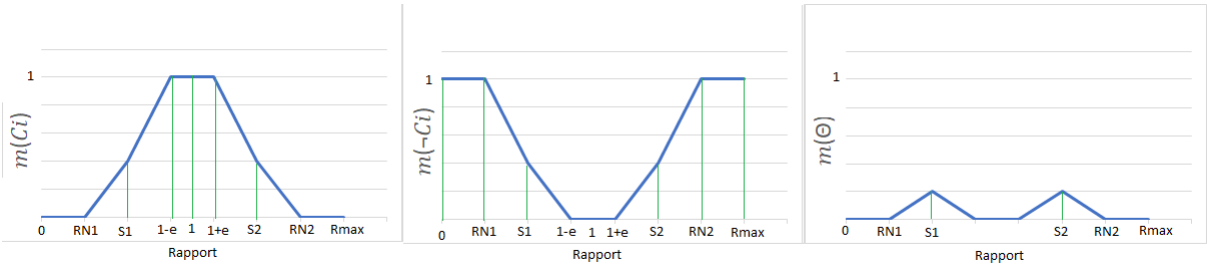


FIGURE 4.23 – Représentation des connaissances pour le critère de la longueur

La première courbe de la figure 4.23 représente la masse de croyance attribuée à l'hypothèse C_i, c'est-à-dire "le candidat C_i est l'homologue de l'objet de référence". Ce qui signifie que si le rapport entre la longueur de l'arête de référence et de l'arête de comparaison est très proche de 1 avec une petite variation e alors la probabilité que cette dernière soit l'objet homologue est égale à 1. Cela signifie également que si le rapport est supérieur à $1 + e$ ou inférieur à $1 - e$ la probabilité que C_i soit l'homologue diminuera en fonction du rapport.

La deuxième courbe représente la masse de croyance attribuée à l'hypothèse $\neg C_i$, qui signifie "le candidat C_i n'est pas l'homologue de l'objet de référence". Lorsque le rapport entre la longueur de l'arête de référence et la longueur de l'arête de comparaison est inférieur à $RN1$ ou supérieur à $RN2$ la probabilité que cette dernière ne représente pas l'objet homologue est égale à 1, cela signifie également que si le rapport est supérieur à $RN1$ ou inférieur à $RN2$ alors cette hypothèse est progressivement improbable.

Finalement, la dernière courbe représente la masse de croyance attribuée à l'ignorance, $m(\Theta)$. Pour le critère de la longueur, l'ignorance est importante lorsque le rapport est au voisinage de $S1$ ou $S2$, c'est-à-dire lorsque la longueur de l'arête de comparaison n'est ni assez longue par rapport à la longueur de l'arête de référence pour conclure avec certitude que ce n'est pas elle, ni proche de 1 pour conclure qu'il s'agit d'arêtes homologues.

4.5 Non fiabilité des sources

Une manière de formuler la fiabilité des sources d'information, et ainsi de diminuer les conflits, consiste à affaiblir les sources en attribuant un coefficient d'affaiblissement pour chaque source d'information. Mais cette solution ne s'applique que si la source non fiable est connue. Sachant que parmi les sources de conflit figurent des sources d'information non fiables, la notion de fiabilité peut être envisagée comme la capacité d'une source à rendre compte de la réalité qu'elle observe avec fidélité.

Nous rappelons ci-dessous la définition de l'opération d'affaiblissement :

$$\begin{cases} \alpha_{m(\mathrm{A})} = (1-\alpha)m(\mathrm{A}), \quad \forall A \subset \Theta \\ \alpha_{m(\Omega)} = (1-\alpha)m(\Theta) + \alpha \end{cases} \tag{4.8}$$

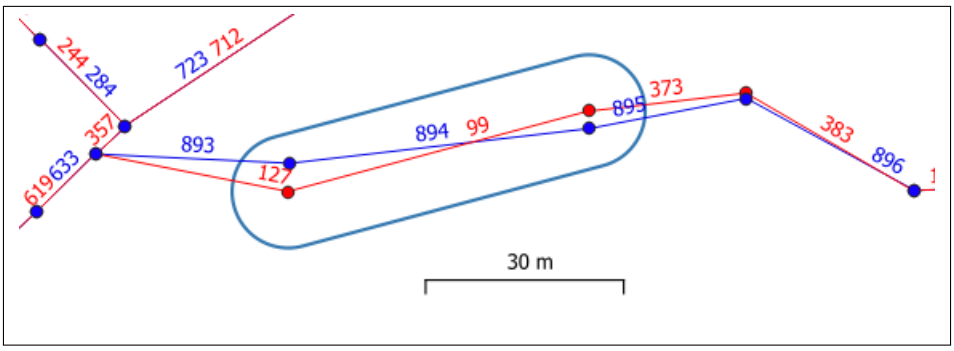


FIGURE 4.24 – Exemple d'affaiblissement

Dans la figure 4.24, Notre objectif est de trouver le correspondant de l'objet de référence numéro 99. Par conséquent, nous commençons par l'étape de sélection des candidats en utilisant un buffer de distance égal à 10 mètres autour de l'arête de référence. Ainsi, à partir de ce buffer, les candidats de l'arrête 99 sont 893, 894 et 895. Dans cet exemple, nous utilisons le critère de distance de Hausdorff pour visualiser les effets de la modification de coefficient d'affaiblissement sur l'attribution de masse pour chaque hypothèse de $S_i = \{C_i, \neg C_i, \Theta\}$.

Dans la figure 4.25, le graphe à gauche montre le jeu de masses attribué aux candidats de l'arête de référence numéro 99 avec un niveau de confiance de $(1-\alpha) = 1$, c'est-à-dire que la source de comparaison est parfaitement fiable. Cela signifie que les masses ne subissent aucune modification après leur initialisation. Au centre, le graphe montre le jeu de masses avec un niveau de confiance $(1-\alpha) = 0.7$, ce qui signifie que la source de comparaison est fiable à 70%. Et le graphe à droite montre le jeu de masses avec un niveau de confiance $(1-\alpha) =$

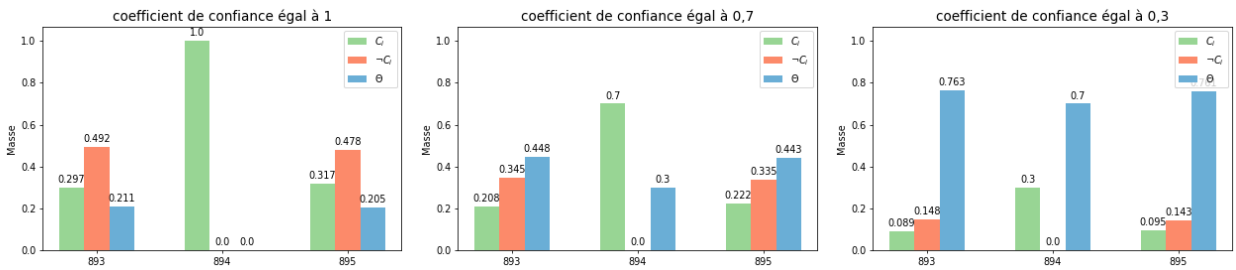


FIGURE 4.25 – Exemple des masses attribuées aux candidats de l'arête numéro 99 en fonction de la distance de hausdorff

0.3, ce qui signifie que la source de comparaison est fiable à 30%. Par conséquent, ces deux derniers graphes seront modifiés en fonction de la formule d'affaiblissement. On remarque dans ces deux derniers graphes que la masse allouée à l'ignorance Θ augmente et que les masses allouées aux autres hypothèses diminuent si le degré d'affaiblissement augmente.

Concernant la valeur du coefficient de confiance utilisé pour la base de données de comparaison des nœuds qui sont détectés à partir d'images à haute résolution [8], la précision de la méthode de détection des ces nœuds est égale à 68.5%, cette valeur a été utilisée comme degré de confiance de la source. Par contre, pour les arêtes, le degré de fiabilité de la source de comparaison a été fixé expérimentalement à 80%.

Chapitre 5

Expérimentations et résultats

5.1 Scénario de l'expérimentation

Dans cette section, nous présentons nos expérimentations et résultats obtenus pour chaque catégorie de formes d'imperfections liées aux réseaux d'assainissement. Nous proposons ainsi le scénario d'expérimentation suivant :

1. Présentation de l'exemple sur lequel nous effectuons l'expérimentation, tous les exemples que nous allons aborder sont détaillés dans la section 4.2.
2. Initialisation des masses de croyance pour chaque hypothèse de l'ensemble $S_i = \{C_i, \neg C_i, \Theta\}$ et pour chaque critère de similarité (la modélisation de ces critères est présentée dans la section 4.4). Nous utilisons les critères de distance euclidienne et de score pour apparier les nœuds et les critères d'angle, de distance de Hausdorff et de longueur pour apparier les arêtes.
3. La fusion de ces critères pour chaque candidat dans le but de synthétiser la connaissance.
4. La fusion de candidats pour faire apparaître l'hypothèse NA, c'est-à-dire "l'objet de référence n'est pas apparié".
5. Le calcul de la probabilité pignistique pour prendre la décision.

5.2 Expérimentations pour les nœuds

Dans le contexte d'appariement des nœuds, nous utilisons les critères de similarité suivants :

- **Distance euclidienne** : elle est calculée à partir des positions des nœuds, ce critère a prouvé son utilité dans plusieurs projets d'appariement de données géographiques. En effet, la position des objets est une connaissance toujours disponible et donc la notion de distance est présente.
- **Score** : il s'agit de la précision de chaque nœud détecté à partir des images à haute résolution.

5.2.1 Appariement d'un nœud : cas d'un seul candidat potentiel

La figure 5.1 montre un exemple de décalage d'un nœud avec un seul candidat. Le nœud de référence numéro 648 ne possède qu'un seul candidat dans la base de données de comparaison, qui est le nœud numéro 270.

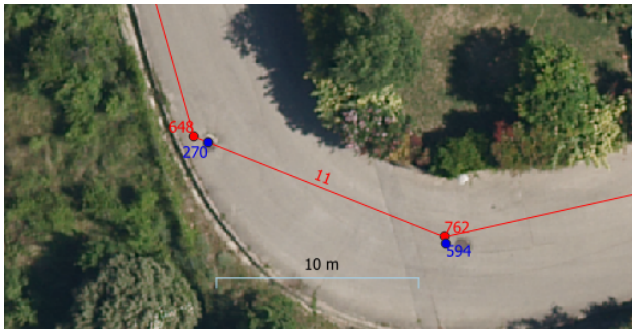


FIGURE 5.1 – Décalage d'un nœud avec un seul candidat

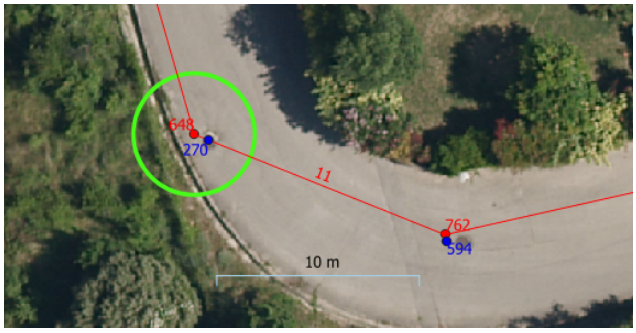


FIGURE 5.2 – Sélection des candidats du nœud numéro 648

Dans la figure 5.2, nous utilisons un buffer de distance égal a 3 mètres pour sélectionner les candidats du nœud numéro 648. le cadre de discernement est défini comme suit :

$$\Theta = \{270, \mathrm{NA}\} \tag{5.1}$$

Dans le jeu de masses présenté dans la figure 5.3, les trois bâtons vert, rouge et bleu représentent respectivement les masses de croyance attribuées par le critère de la distance euclidienne et par le critère du score, aux trois hypothèses $\{C_i, \neg C_i, \Theta\}$ avec un degré de confiance de la source égal à 68%.

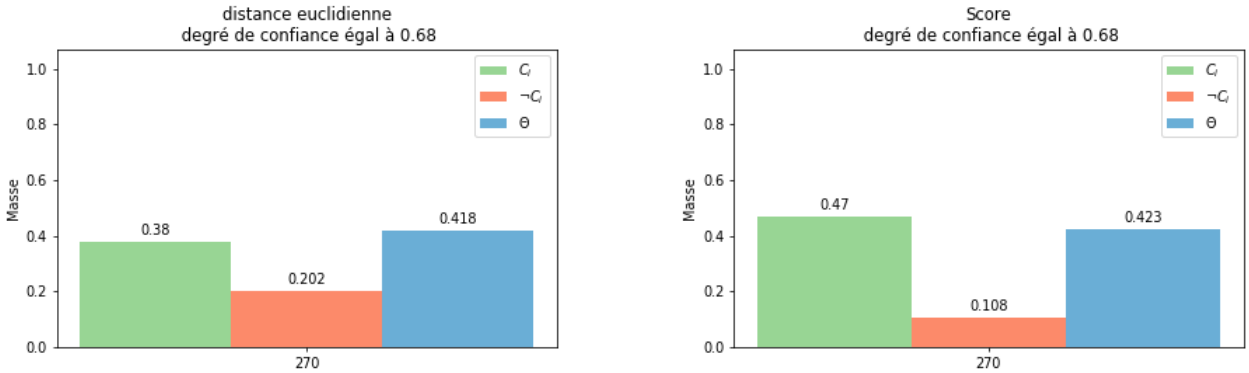


FIGURE 5.3 – Initialisation des masses de croyance pour chaque critère

La figure 5.4 illustre le jeu de masses résultant de la fusion des critères. En analysant la Figure 5.4, nous constatons qu'il existe un conflit de 13,6% entre les mesures. Par contre, nous

pouvons exprimer avec 53,5% de certitude que le nœud numéro 270 est le bon homologue.

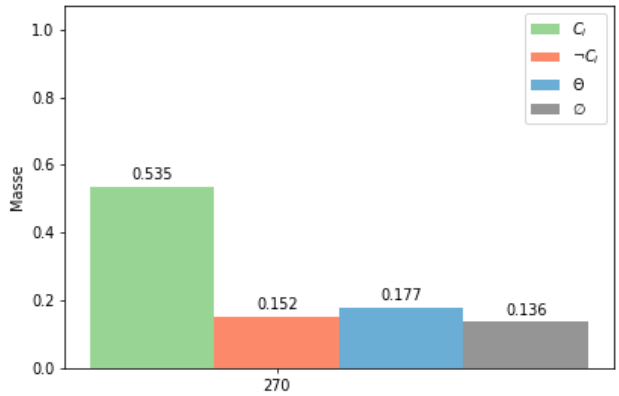


FIGURE 5.4 – Jeu de masses après la combinaison des mesures

La figure 5.5 montre le jeu de masses après la normalisation de Dempster (voir la section 2.3.3). En analysant cette figure, on remarque que l'hypothèse NA apparaît, et la masse attribuée à celle-ci est le résultat de la fusion de toutes les masses de croyance attribuées aux hypothèses $\neg C_i$, pour $i = 1 \ldots N$, mais dans ce cas il n'y a qu'un seul candidat, ce qui signifie que $m(\text{NA}) = m(\neg C_i)$.

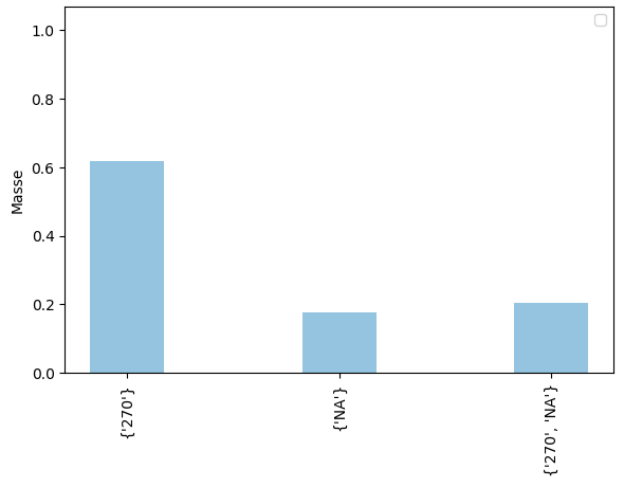


FIGURE 5.5 – Jeux de masse après la normalisation

Après avoir calculé la probabilité pignistique (voir figure 5.6), l'hypothèse 270 est choisie, parce que la probabilité pignistique attribuée a la valeur maximale. **Par conséquent, le nœud de référence numéro 648 est apparié avec le nœud de comparaison 270.**

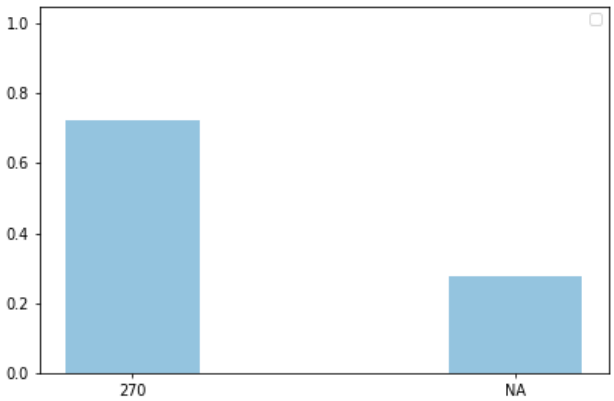


FIGURE 5.6 – Probabilité pignistique pour chaque hypothèse

5.2.2 Appariement d'un nœud : cas de plusieurs candidats potentiels

La figure 5.7 présente un exemple de décalage d'un nœud ayant plusieurs candidats pour l'appariement.

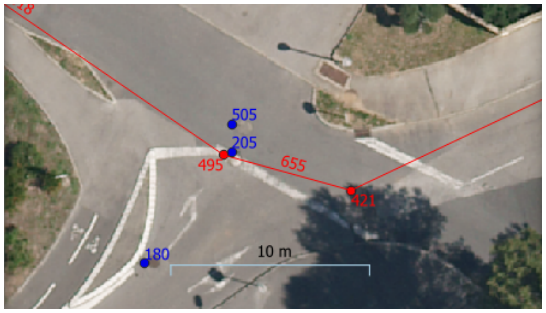


FIGURE 5.7 – Décalage d'un nœud avec plusieurs candidats

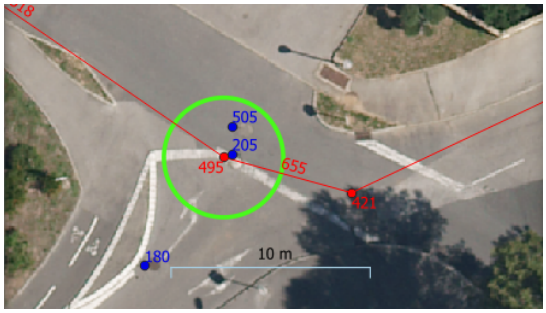


FIGURE 5.8 – Sélection des candidats du nœud numéro 495

Dans la figure 5.8, nous utilisons un "buffer" (zone tampon) de distance égale à 3 mètres pour sélectionner les candidats du nœud numéro 495, qui possède plusieurs candidats (les nœuds de comparaison numéro 205 et 505). Le cadre de discernement est défini comme suit :

$$\Theta = \{205, 505, \mathrm{NA}\} \tag{5.2}$$

La figure 5.9 présente les jeux de masses pour chaque candidat attribuées par le critère de la distance euclidienne et par le critère du score aux trois hypothèses $\{C_i, \neg C_i, \Theta\}$ avec un degré de confiance de la source égal à 68%. On peut donc affirmer en termes de distance euclidienne, avec certitude, que le candidat 505 n'est pas l'objet homologue de l'objet de référence. Par contre, nous constatons que les candidats 205 et 505 ont des masses de croyance très proches selon le score.

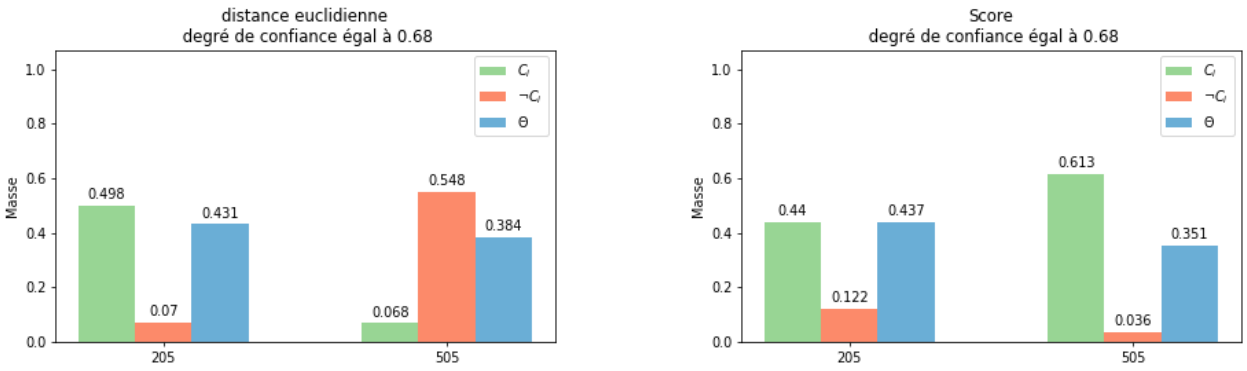


FIGURE 5.9 – Initialisation des masses de croyance pour chaque critère

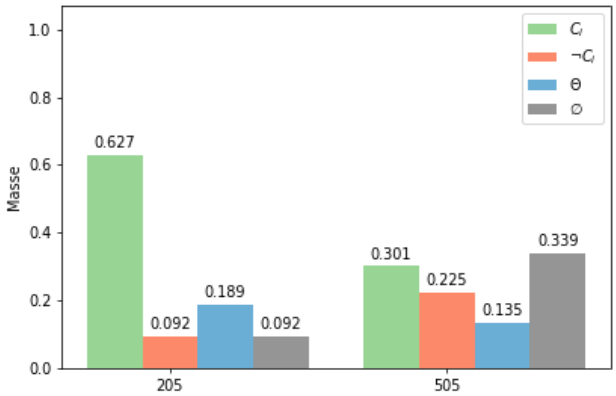


FIGURE 5.10 – Jeu de masses après la combinaison des mesures

La figure 5.10 illustre le jeu de masses résultant de la fusion des critères. En analysant ces derniers, nous trouvons qu'il existe un conflit de 33,9% entre les mesures pour le candidat 505. Par contre, pour le candidat 205, le conflit entre les deux mesures est de 0,92%. Ce qui signifie que nous pouvons exprimer avec 62,7% de certitude que le nœud numéro 205 est le nœud homologue.

Si nous effectuons une fusion des candidats, toujours en utilisant l'opérateur conjonctif sans normalisation, nous obtenons les jeux de masses illustrés sur la Figure 5.11 à gauche. Nous remarquons qu'après la fusion des candidats, la masse de croyance associée au conflit 58.8%, est assez importante, par rapport aux autres valeurs obtenues. Ce conflit est dû au fait que différents critères soutiennent plusieurs candidats en même temps.

La masse de croyance attribuée à l'hypothèse NA résulte de la fusion de toutes les masses de croyance attribuées aux hypothèses $\neg C_i$, Elle est calculée de la manière suivante :

$$m(\mathrm{NA}) = m(\neg C_{205}) * m(\neg C_{505}) \tag{5.3}$$

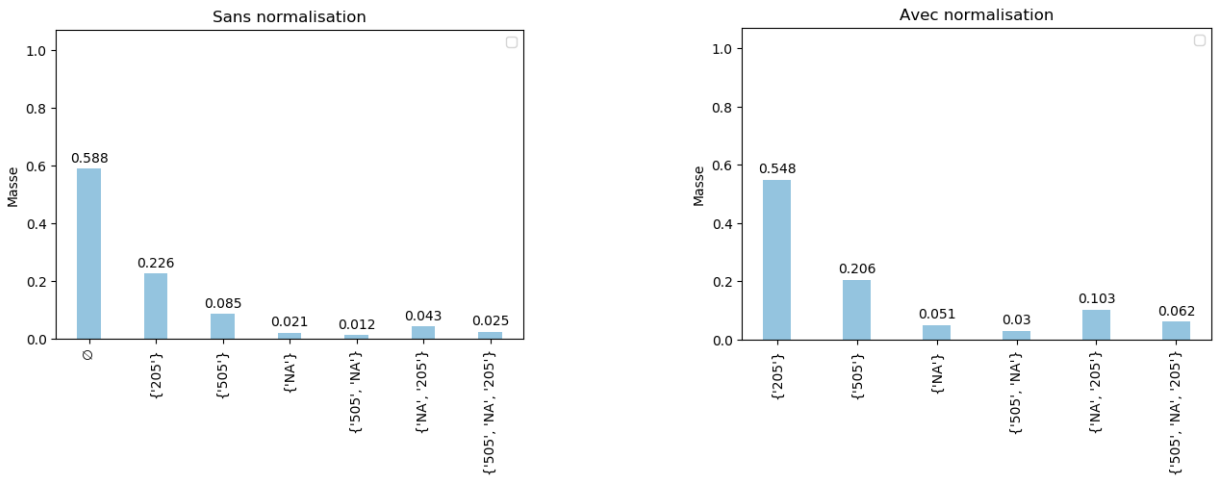


FIGURE 5.11 – Fusion des candidats

Après avoir normalisé les masses grâce à l'opérateur de Dempster, c'est-à-dire que toutes les masses de croyance sont normalisées par la masse de croyance allouée au conflit, on remarque sur la figure 5.11 à droite, que l'hypothèse liée au candidat 205 est clairement plus éloignée de celle attribuée au candidat 505, avec une masse de croyance égale à 0.548.

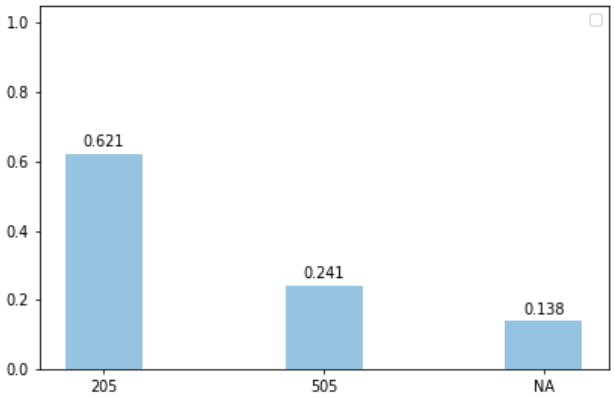


FIGURE 5.12 – Probabilité pignistique pour chaque hypothèse du cadre de discernement

Au terme du calcul de la probabilité pignistique (voir figure 5.12), l'hypothèse 205 est choisie, puisque la probabilité pignistique attribuée a la valeur maximale, P(205)=0,62. **Par conséquent, le nœud de référence numéro 495 est apparié avec le nœud candidat numéro 205.**

5.3 Expérimentations pour les arêtes

Dans le cadre d'appariement des arêtes, nous utilisons les critères de similarité suivants :

- **Angle** : ce critère est utilisé dans plusieurs travaux [51, 36, 63]. En général, la différence entre ces derniers réside dans la manière avec laquelle l'angle est mesuré : choix des axes, des points à utiliser, etc. Notre méthode de calcul des angles est présentée dans la section 4.4.2.3.

- **Distance de Hausdorff** : cette distance est définie comme la distance maximale d'un ensemble par rapport à l'objet le plus proche dans l'ensemble cible (voir section 3.2.1.1).

- **Longueur** : ce critère a prouvé son utilité dans plusieurs projets d'appariement géographique, parce que dans ce cadre la longueur des arêtes est une connaissance toujours disponible.

5.3.1 Décalage uniforme de type 1 : 1

Dans la figure 5.13, nous présentons un exemple de décalage uniforme de type 1 : 1.

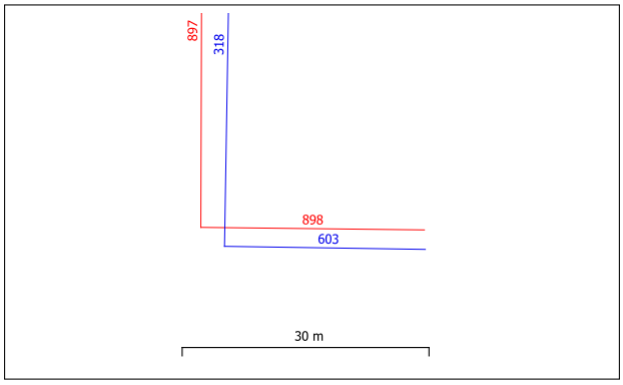


FIGURE 5.13 – Exemple de décalage uniforme avec une cardinalité de 1 : 1

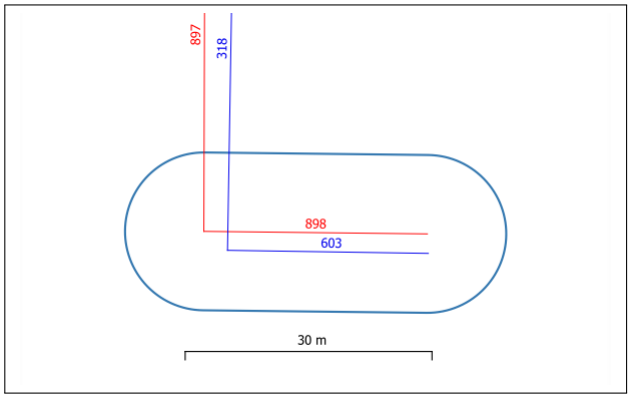


FIGURE 5.14 – Sélection des candidats de l'arête numéro 898

La figure 5.14 montre un buffer de distance égale à 10 mètres pour sélectionner les candidats de l'arête numéro 898. Cette dernière possède plusieurs candidats dans la base de données de comparaison, qui sont les arêtes numéro 318 et 603. Donc le cadre de discernement est défini comme suit :

$$\Theta = \{318, 603, \mathrm{NA}\} \tag{5.4}$$

La figure 5.15 présente les jeux de masses pour chaque candidat attribuées par le critère de l'angle, distance de Hausdorff et la longueur, aux trois hypothèses $\{C_i, \neg C_i, \Theta\}$ avec un degré de confiance de la source égal à 80%. Grâce à l'analyse du jeu de masses présenté dans la figure 5.15, nous pouvons dire avec certitude que l'objet de référence ne peut pas être apparié

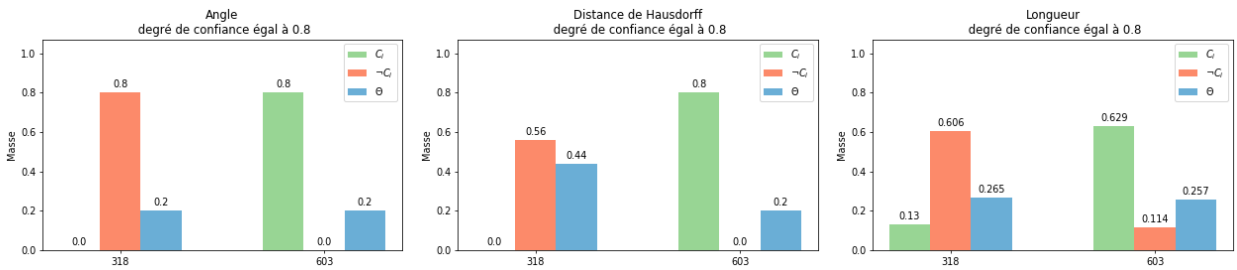


FIGURE 5.15 – Initialisation des masses de croyance pour chaque critère

avec le candidat numéro 318 du point de vue de l'angle et de la distance de Hausdorff, mais nous pouvons affirmer avec certitude que le candidat numéro 603 est l'homologue de l'arête de référence du point de vue des trois critères.

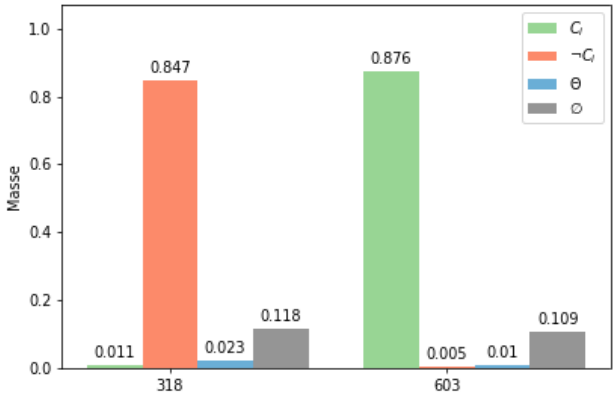


FIGURE 5.16 – Jeu de masses après la combinaison des mesures

La figure 5.16 illustre le jeu de masses résultant de la fusion des critères. En analysant ces derniers. Nous pouvons affirmer avec 87.6% de certitude que l'homologue est 603 et avec un peu de conflit égal à 10.9%. Par contre nous remarquons que le candidat numéro 318 n'est pas l'homologue avec 84.7% de certitude et un conflit de 11.8%.

Maintenant nous effectuons une fusion des candidats, toujours en utilisant l'opérateur conjonctif sans normalisation, nous obtenons les jeux de masses illustrés sur la Figure 5.17 à gauche. Nous constatons qu'après la fusion des candidats, la masse de croyance associée au conflit est égale à 22.4%, mais que la masse la plus importante est allouée à l'hypothèse 603 avec une valeur égale à 76.2%, les autres hypothèses ont une masse négligeable.

Après la normalisation des masses grâce à l'opérateur de Dempster, on remarque sur la figure 5.17 à droite que l'hypothèse liée au candidat 603 est clairement plus éloignée par rapport aux autres hypothèses, avec une masse de croyance égale à 98.3%.

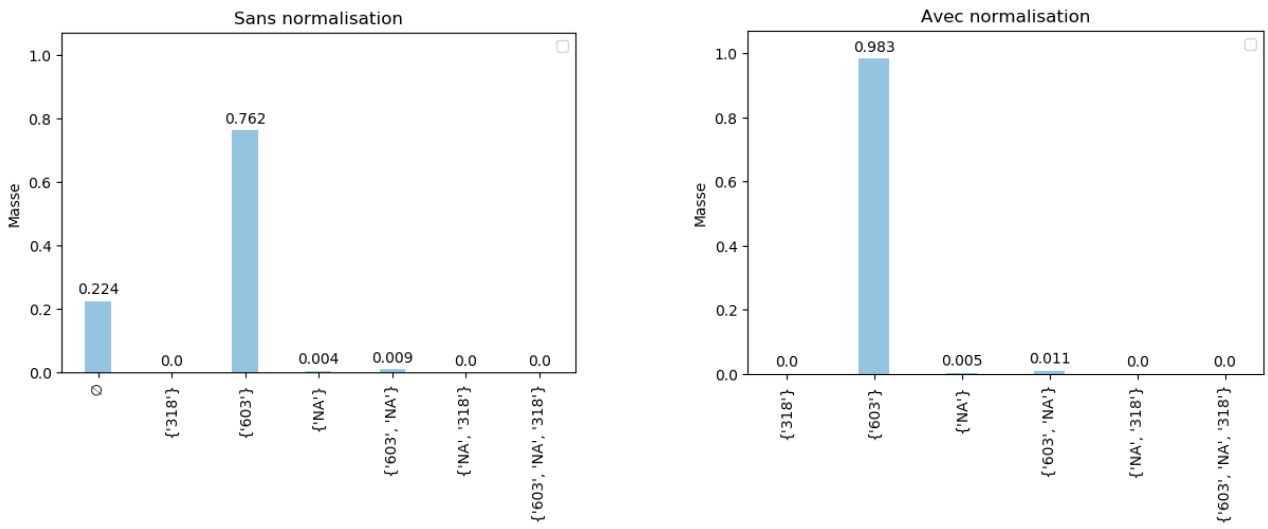


FIGURE 5.17 – Fusion des candidats

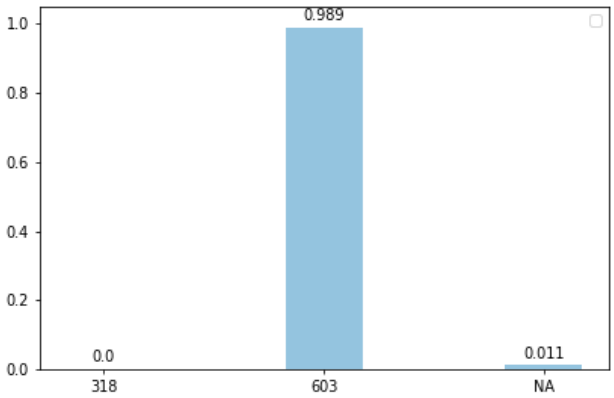


FIGURE 5.18 – Probabilité pignistique pour chaque hypothèse du cadre de discernement

A la fin du calcul de la figure de probabilité pignistique 5.18, l'hypothèse 603 est choisie, puisque la probabilité pignistique attribuée à une valeur maximale, P(603)=0.989. **Par conséquent, l'arête de référence numéro 898 est appariée avec l'arête candidate numéro 603.**

5.3.2 Décalage non uniforme de type 1 : 1

La figure 5.19 présente l'exemple de décalage non uniforme entre les lignes avec une cardinalité de 1 : 1.

La figure 5.20 montre un buffer d'une distance égale à 10 mètres pour sélectionner les candidats de l'arête numéro 99. Cette dernière possède plusieurs candidats dans la base de données de comparaison, qui sont les arêtes numéro 893, 894 et 895. Donc le cadre de discer-

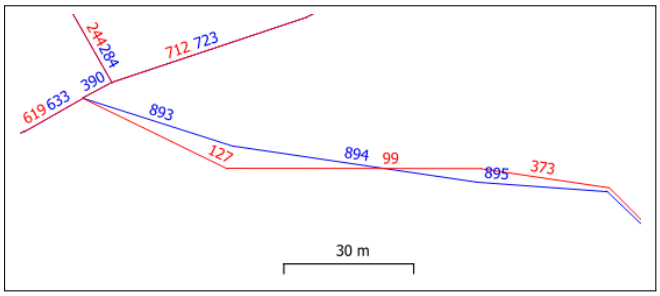


FIGURE 5.19 – Exemple de décalage non uniforme avec une cardinalité de 1 : 1

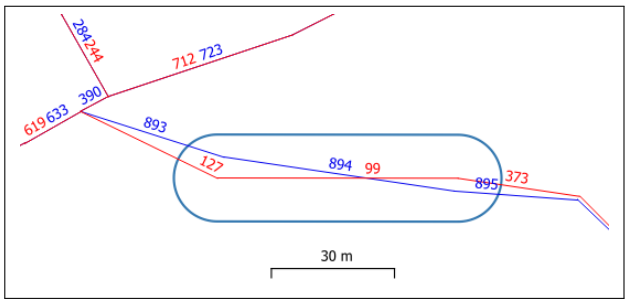


FIGURE 5.20 – Sélection des candidats de l'arête numéro 99

nement est défini comme suit :

$$\Theta = \{893, 894, 895, \mathrm{NA}\} \tag{5.5}$$

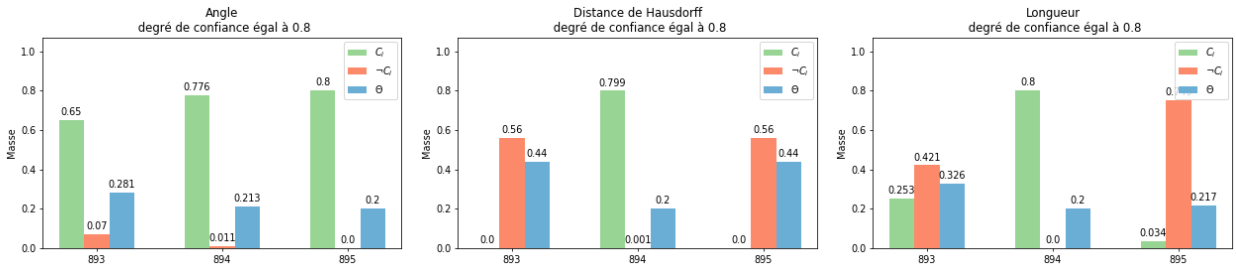


FIGURE 5.21 – Initialisation des masses de croyance pour chaque critère

La figure 5.21 montre les jeux de masses pour chaque candidat attribués par le critère de l'angle, Distance de Hausdorff et la longueur, aux trois hypothèses $\{C_i, \neg C_i, \Theta\}$ avec un degré de confiance de la source égal à 80%. En analysant ces figures, nous pouvons dire avec certitude que l'objet de référence ne peut pas être apparié avec le candidat numéro 895 du point de vue de la distance de Hausdorff et de la longueur, et ne peut pas être apparié avec le candidat 893 en termes de la distance de Hausdorff. Mais nous pouvons affirmer avec certitude que le candidat numéro 894 est l'homologue de l'arête de référence du point de vue des trois critères.

La figure 5.22 illustre le jeu de masses résultant de la fusion des critères. En analysant ce dernier, nous pouvons affirmer avec 98% de certitude que l'homologue est le candidat numéro 864 avec une valeur de conflit égale à 1.1% seulement. D'autre part, nous remarquons qu'il y a un conflit égal à 71,5% pour le candidat numéro 895, parce que ce candidat est l'homologue de l'arête de référence selon le point de vue de l'angle, mais pas selon les autres critères.

Après avoir fusionné les candidats, nous obtenons les jeux de masses illustrés dans la

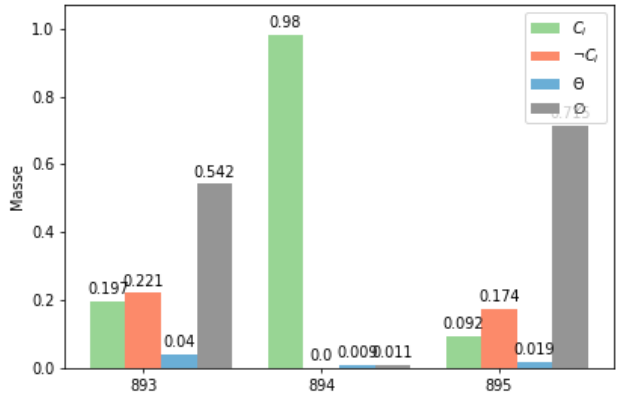


FIGURE 5.22 – Jeu de masses après la combinaison des mesures

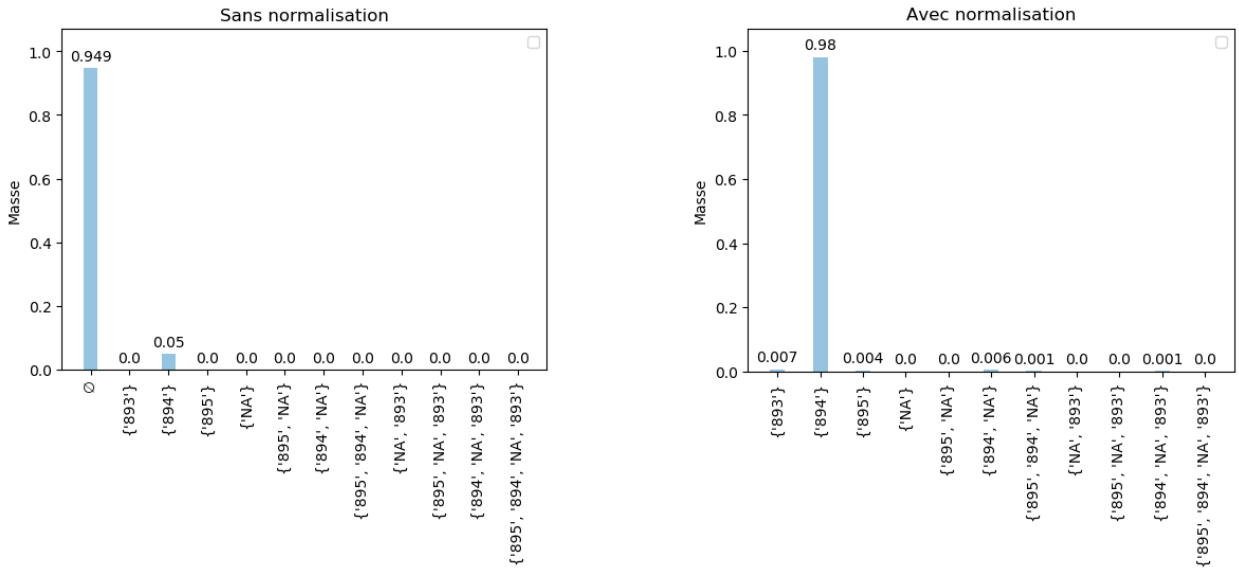


FIGURE 5.23 – Fusion des candidats

Figure 5.23 à gauche. Nous remarquons qu'après la fusion des candidats, la masse de croyance associée au conflit est égale à 94.9% parce que différents critères soutiennent plusieurs candidats en même temps.

Après la normalisation des masses grâce à l'opérateur de Dempster, nous constatons sur la figure 5.23 à droite que l'hypothèse liée au candidat 894 est plus éloignée de toutes les autres hypothèses, avec une masse de croyance égale à 0.98.

Au terme du calcul de la probabilité pignistique figure 5.24, l'hypothèse 894 est choisie, puisque la probabilité pignistique a atteint une valeur maximale, P(894)=0.984. **Par conséquent, l'arête de référence numéro 99 est appariée avec l'arête candidate numéro 894.**

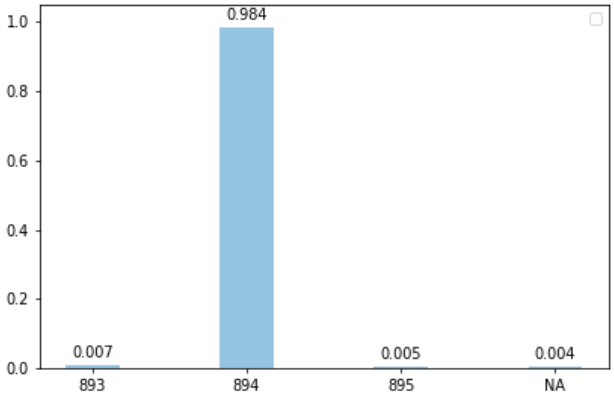


FIGURE 5.24 – Probabilité pignistique pour chaque hypothèse du cadre de discernement

5.3.3 Décalage uniforme de type 1 : n

La figure 5.25 montre l'exemple de décalage uniforme entre les lignes avec une cardinalité de 1 : n que nous aborderons dans cette section.

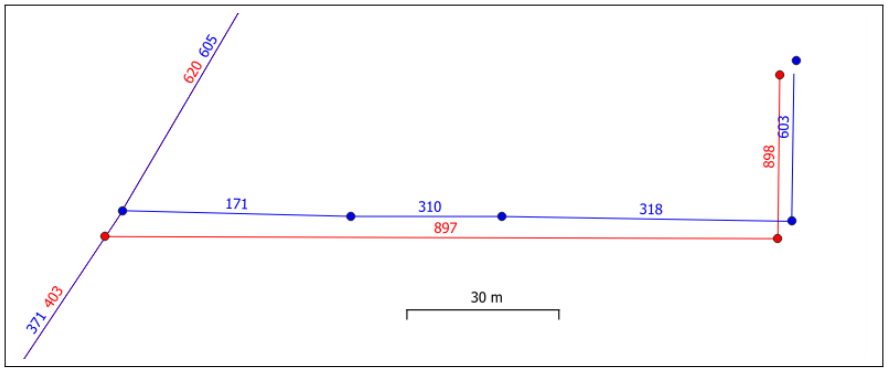


FIGURE 5.25 – Décalage uniforme avec une cardinalité de 1 : n

Après l'application d'un buffer de distance de 10 mètres pour sélectionner les candidats de l'arête de référence numéro 897 (voir figure 5.26), cette dernière possède plusieurs candidats dans la base de données de comparaison, qui sont les d'arêtes numéro 171, 310, 318, 371, 605 et 603. De ce fait, le cadre de discernement est défini comme suit :

$$\Theta = \{171, 310, 318, 371, 605, 603, \mathrm{NA}\} \tag{5.6}$$

La figure 5.27 illustre les jeux de masses pour chaque candidat attribuées par le critère d'angle, de distance de Hausdorff et de longueur, aux trois hypothèses $\{C_i, \neg C_i, \Theta\}$ avec un niveau de confiance de la source égal à 80%. Grâce à l'analyse de ces chiffres, on peut affirmer avec certitude que l'objet de référence ne peut être apparié avec aucun candidat en termes de

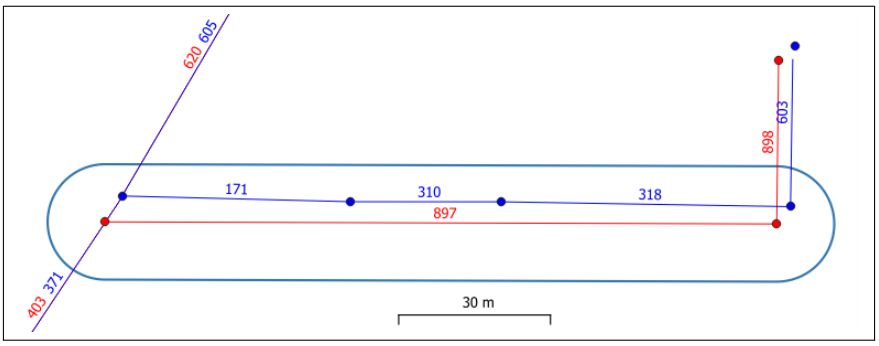


FIGURE 5.26 – Sélection des candidats de l'arête numéro 897

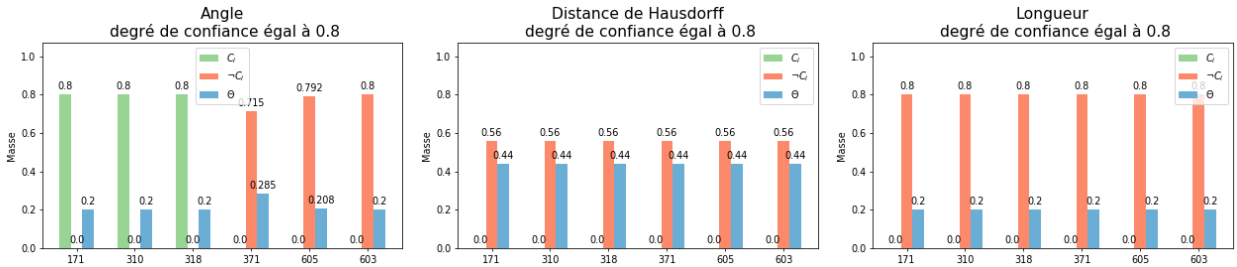


FIGURE 5.27 – Initialisation des masses de croyance pour chaque critère

distance de Hausdorff et de longueur et ne peut être apparié avec les candidats numéro 371, 605 et 603 en termes de distance de l'angle. Mais également du point de vue de l'angle, l'arête de référence peut être appariée avec les candidats 171, 310 et 318.

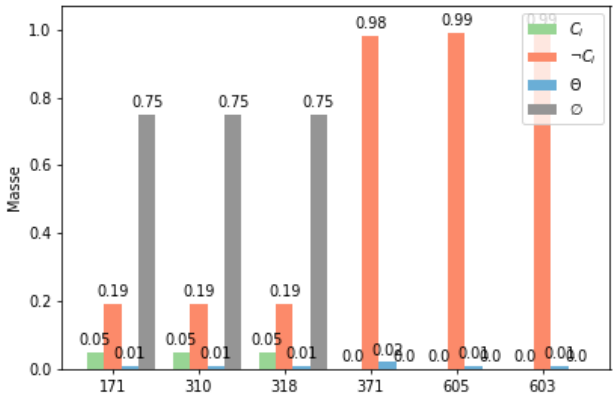


FIGURE 5.28 – Jeu de masses après la combinaison des mesures

La figure 5.28 montre le jeu de masses résultant de la fusion des critères. Après avoir analysé celui-ci, nous pouvons affirmer avec certitude que les candidats numéro 371, 605 et 603 ne sont pas les homologues de l'arête de référence. Par contre, nous remarquons qu'il y a un conflit de 75% pour les candidats numéro 171, 310 et 318, car selon le point de vue de l'angle, ces candidats sont les homologues de l'arête de référence et selon les autres critères ne sont pas les homologues.

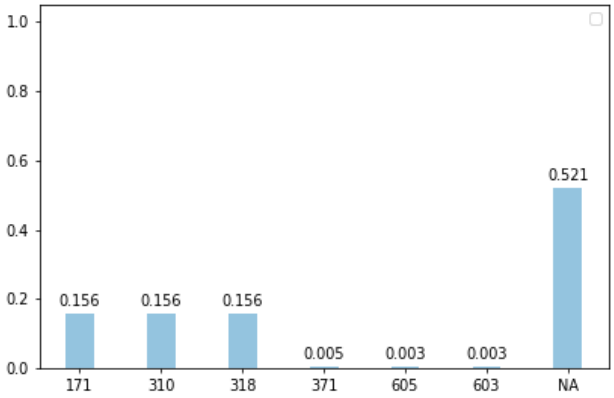


FIGURE 5.29 – Probabilité pignistique pour chaque hypothèse du cadre de discernement

Après la fusion des candidats et la normalisation du jeu de masses obtenu avec l'opérateur de Dempster, nous présentons la probabilité pignistique dans la figure 5.29 pour chaque hypothèse du cadre de discernement. Ce qui signifie que l'hypothèse NA est choisie, parce que la probabilité pignistique a atteint une valeur maximale, $P(\mathrm{NA}) = 0.521$. **Par conséquent, l'arête de référence numéro 897 ne peut être appariée avec aucune candidate.**

Problématique

Les résultats de l'exemple précédent montrent que l'objet de référence ne peut être apparié avec aucun candidat. Mais en analysant cet exemple, on constate que l'objet de référence doit être apparié avec les trois candidats 171, 310 et 318 en même temps. Cela signifie une cardinalité de 1 : n. Donc pour les apparier, nous utilisons le principe de "bonne continuité".

Solution

Le principe de « Bonne continuité » est souvent utilisé comme guide pour établir des règles de connexion géométrique. Dans le cadre des réseaux d'assainissement ce principe est utilisé pour connecter les arêtes de la manière suivante : deux arêtes peuvent être connectées à condition que leur angle de déflexion soit inférieur à un seuil déterminé. L'angle de déflexion, qui va de 0° à 180°, représente le degré de déviation formé par deux segments de route reliés [62], à partir de ce moment, on appelle une bonne continuité des arêtes des réseaux d'assainissement un stroke.

la figure 5.30 à gauche présente les données spatiales des réseaux d'assainissement or-

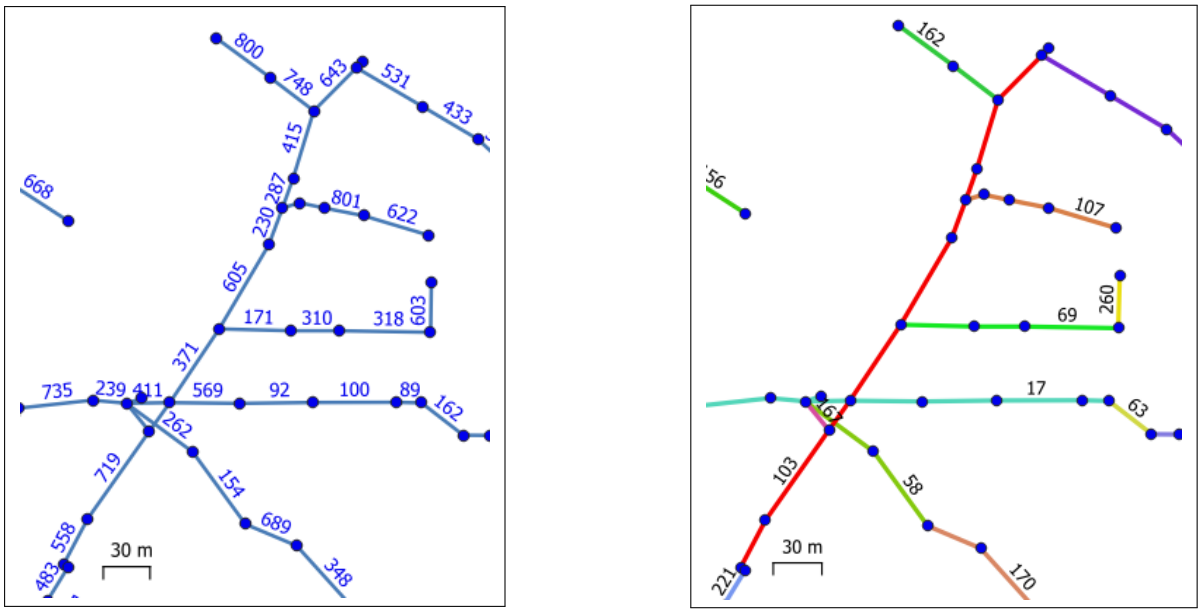


FIGURE 5.30 – Exemple des strokes

ganisées de la manière précédente "nœud-arête", et la figure à droite montre les données traitées par connexion de strokes. Il y a plusieurs strokes. Par exemple : le stroke rouge numéro 103<558, 719, 767, 371, 605, 230, 287, 415, 643>, le stroke vert numéro 69<171, 310, 318>, le stroke jaune numéro 260 <603>. "Une bonne continuation" fait généralement référence à un angle approximativement égal (180°) entre les segments. La morphologie géométrique d'un stroke est caractérisée par une distribution lisse et continue dans l'espace, ce qui maintient l'intégrité des canalisations souterraines urbaines. Par conséquent, le stroke peut être utilisé comme unité de base pour l'appariement des arêtes des réseaux d'assainissement.

Après avoir détecté les strokes de l'exemple précédent de décalage uniforme entre les lignes de type 1 : 1, on obtient les résultats illustrés dans la figure 5.31.

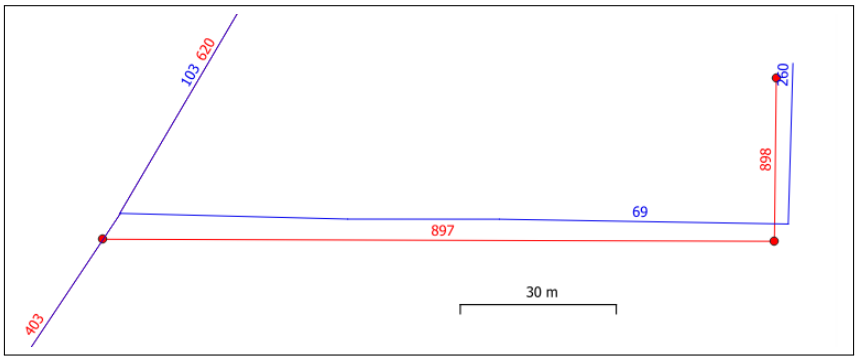


FIGURE 5.31 – Décalage uniforme avec une cardinalité de 1 : n

En utilisant un buffer de distance de 10 mètres pour sélectionner les candidats du stroke

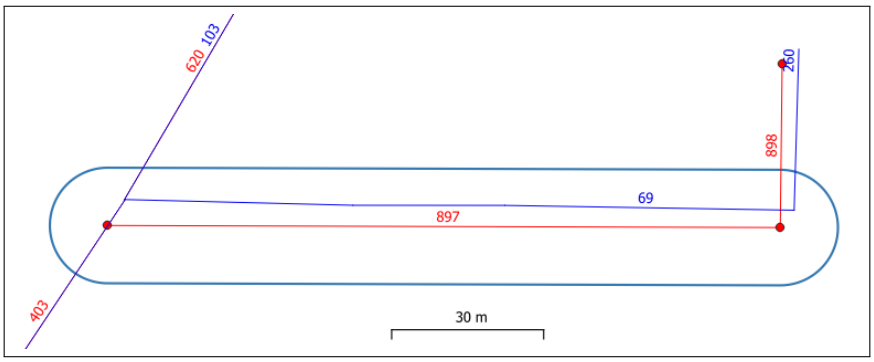


FIGURE 5.32 – Détection des strokes

de référence numéro 897 (voir figure 5.32), il se trouve que ce dernier a plusieurs candidats dans la base de données de comparaison, qui sont les strokes numéro 69, 260 et 103. Par conséquent, le cadre de discernement est défini comme suit :

$$\Theta = \{69, 260, 103, \mathrm{NA}\} \tag{5.7}$$

FIGURE 5.33 – Initialisation des masses de croyance pour chaque critère

La figure 5.33 illustre les jeux de masses pour chaque candidat attribuées par le critère de distance de Hausdorff et de longueur, aux trois hypothèses $\{C_i, \neg C_i, \Theta\}$ avec un niveau de confiance de la source égal à 80%. Grâce à l'analyse de ces chiffres, on peut affirmer avec certitude que l'objet de référence ne peut être apparié avec les strokes numéro 103 et 260 en termes de distance de Hausdorff et de longueur. Par contre, nous pouvons affirmer avec certitude que l'homologue du stroke de référence est le candidat numéro 69 en fonction des deux critères.

La figure 5.34 montre le jeu de masses résultant de la fusion des critères. Après avoir analysé celles-ci, nous pouvons affirmer avec 91% de certitude que les candidats numéro 103 et 260 ne sont pas les homologues du stroke de référence mais que le stroke numéro 69 est l'homologue avec une certitude de 95%.

Lorsque la probabilité pignistique est calculée (figure 5.35), le stroke numéro 69 est choisie, parce que la probabilité pignistique a atteint une valeur maximale, P(69)=0.977. **En consé-**

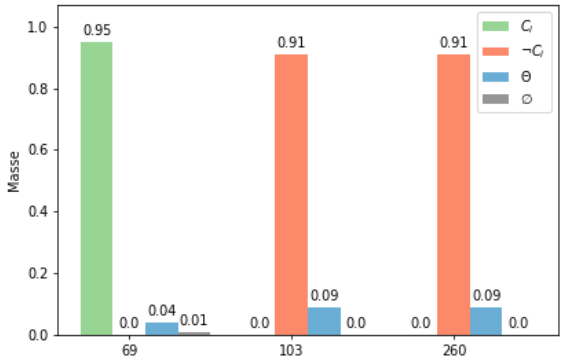


FIGURE 5.34 – Jeu de masses après la combinaison des mesures

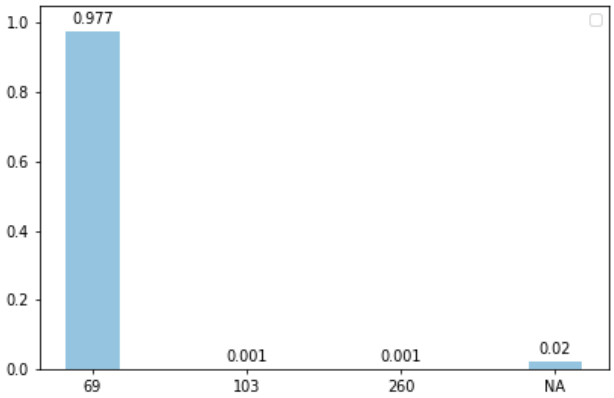


FIGURE 5.35 – Probabilité pignistique pour chaque hypothèse du cadre de discernement

quence, le stroke de référence numéro 897 est apparié avec le stroke candidat numéro 69.

5.3.4 Décalage non uniforme de type 1 : n

La figure 5.36 présente l'exemple de décalage non uniforme entre les lignes avec une cardinalité de 1 : n que nous traitons dans cette section.

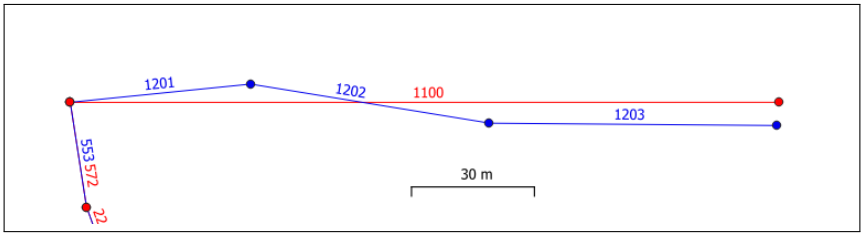


FIGURE 5.36 – Décalage non uniforme avec une cardinalité de 1 : n

Après avoir détecté les strokes de l'exemple précédent de décalage non uniforme entre les lignes de type 1 : n, on obtient les résultats illustrés dans la figure 5.37. En analysant cette figure, on remarque que le stroke principal numéro 2122 est composé de trois arêtes numérotées

1201, 1202 et 1203 dans la figure 5.36.

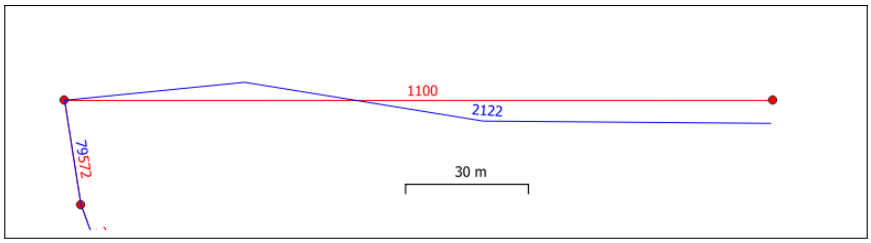


FIGURE 5.37 – Détection des strokes

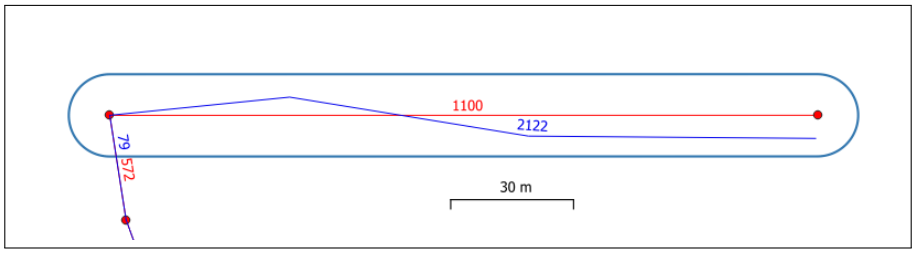


FIGURE 5.38 – Sélection des candidats du stroke numéro 1100

En appliquant un buffer de distance de 10 mètres pour sélectionner les candidats du stroke de référence numéro 1100 (voir la figure 5.38), il se trouve que ce dernier a deux candidats dans la base de données de comparaison, qui sont les strokes numéro 2122 et 79. Par conséquent, le cadre de discernement est défini comme suit :

$$\Theta = \{2122, 79, \mathrm{NA}\} \tag{5.8}$$

FIGURE 5.39 – Initialisation des masses de croyance pour chaque critère

La figure 5.39, illustre les jeux de masses pour chaque candidat attribuées par le critère de distance de Hausdorff et de longueur, aux trois hypothèses $\{C_i, \neg C_i, \Theta\}$ avec un niveau de confiance de la source égal à 80%. Grâce à l'analyse de ces chiffres, on peut affirmer avec certitude que l'objet de référence ne peut être apparié avec le stroke numéro 79 en termes de distance de Hausdorff et de longueur. Par contre, nous pouvons affirmer avec certitude que l'homologue du stroke de référence est le candidat numéro 2122 en fonction des deux critères.

La figure 5.40 présente le jeu de masses résultant de la fusion des critères. En analysant

ce dernier, nous pouvons affirmer avec 95% de certitude que le candidat numéro 2122 est l'homologue du stroke de référence et que le stroke numéro 79 n'est pas l'homologue avec une certitude de 91%.

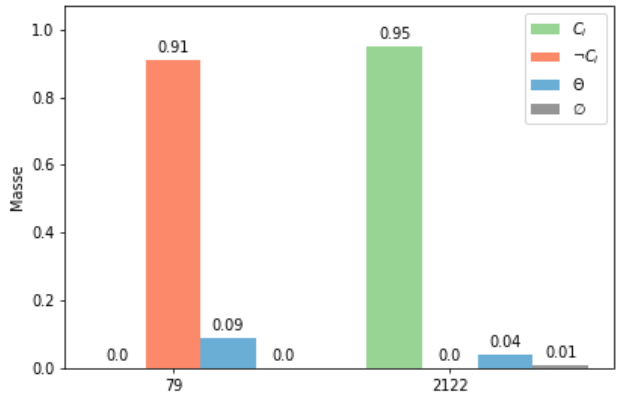


FIGURE 5.40 – Jeu de masses après la combinaison des mesures

En calculant la probabilité pignistique (figure 5.41), le stroke numéro 2122 est choisi, car la probabilité pignistique a atteint une valeur maximale, P(2122)=0.977. **Par conséquent, le stroke de référence numéro 1100 est apparié avec le stroke de référence numéro 2122.**

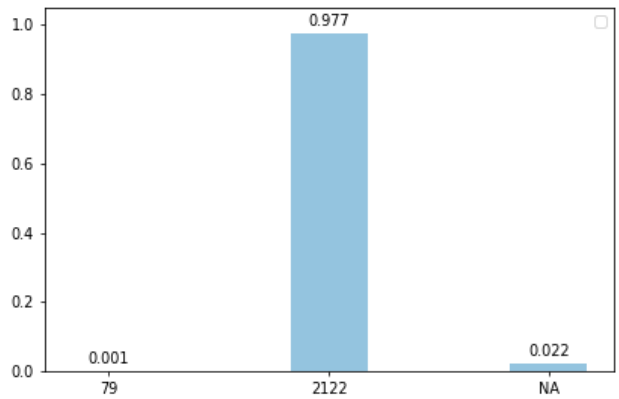


FIGURE 5.41 – Probabilité pignistique pour chaque hypothèse du cadre de discernement

5.4 Synthèse et discussion

Rappelons que la fusion des critères et des candidats a été effectuée à l'aide de l'opérateur conjonctif de Dempster. Cet opérateur a été critiqué en raison de l'absence de redistribution du conflit mais de son utilisation pour normaliser l'ensemble final des masses. La faiblesse de cet opérateur est principalement liée à la sensibilité du facteur de normalisation $[1-\mathrm{m}(\phi)]^{-1}$

lorsque la valeur du conflit est élevée [58].

En partant de ce constat, [7] a analysé la sensibilité de l'opérateur de Dempster au conflit. Sur la courbe de variation du coefficient de normalisation en fonction du conflit présentée dans [7], nous remarquons que le facteur de normalisation devient très sensible lorsque la valeur du conflit est supérieure à 0,95, autrement dit, une petite variation du conflit entraîne une grande variation du facteur de normalisation (voir Figure 5.42).

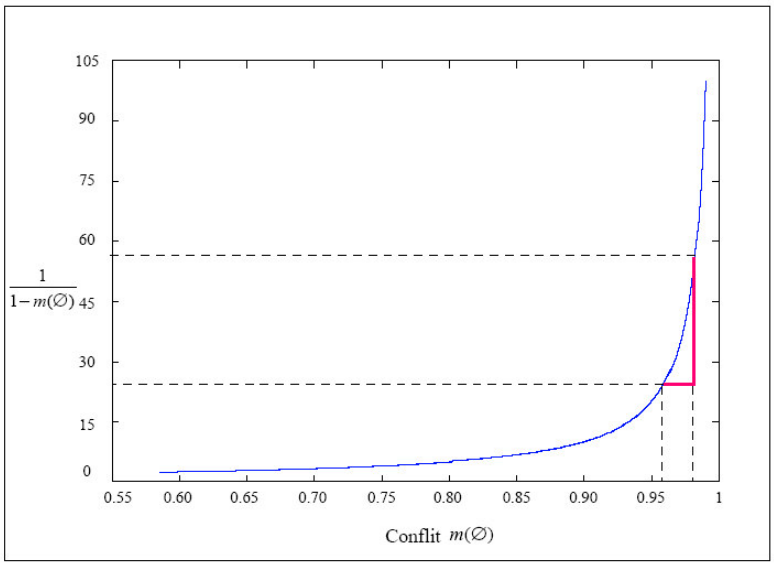


FIGURE 5.42 – Variation du coefficient de normalisation en fonction du conflit [7]

Cependant, nous avons choisi d'utiliser l'opérateur de Dempster pour nos expériences, en premier lieu pour ses propriétés (commutativité, associativité), pour la simplicité de la méthode, et pour le fait que nous ne savons pas quel est le meilleur ordre dans lequel considérer les critères.

Conclusions et perspectives

Dans le contexte de la fusion de bases de données géographiques, nous avons abordé la problématique de l'appariement de données spatiales imparfaites des réseaux d'assainissement.

Pour cela, nous avons présenté un état d'art sur la gestion des données imparfaites dans lequel nous traitons les différents formalismes de la gestion de ces données. Puis, nous avons passé en revue certains travaux dans la littérature qui utilisent la théorie des fonctions de croyance pour la gestion des données imparfaites. Nous avons présenté par la suite les outils mathématiques de la théorie des fonctions de croyance.

Notre contribution concerne, en premier lieu, la catégorisation des différentes formes d'imperfections liées aux nœuds et aux arêtes des réseaux d'assainissement, qui nous a permis de traiter distinctement chaque catégorie. En deuxième lieu, l'adaptation du processus d'appariement géographique proposée par [29] en utilisant la théorie des fonctions de croyance, nous a permis de prendre en compte les imperfections présentes dans les données des réseaux d'assainissement. Dans cette dernière étape nous avons modélisé les différentes mesures de similarité utilisées pour apparier les données des réseaux d'assainissement.

Différentes perspectives d'amélioration peuvent être définies :

- L'automatisation des seuils utilisés pour définir les critères d'appariement : dans nos tests expérimentaux, les seuils de sélection des candidats et ceux utilisés dans les critères d'appariement ont été déterminés de manière expérimentale. Pour améliorer cet aspect, on peut fixer les seuils utilisés par les experts du domaine.
- Une meilleure initialisation des masses de croyance : afin d'améliorer l'étape d'initialisation des masses de croyance, nous pouvons définir un nombre de n critères au départ. Puis nous pouvons comparer les résultats obtenus avec l'ensemble de toutes les

combinaisons possibles, c'est-à-dire 2 critères, 3 critères,... n critères. De cette façon, l'ensemble des masses de croyance peut être affiné et nous pouvons faire une analyse statistique sur l'importance des critères.

Références

[1] M FOULLOY Laurent Professeur À POLYTECH. « Fonctions de Croyance : de la théorie à la pratique ». Thèse de doct. Université d'Artois, 2012.

[2] Alain APPRIOU. « Probabilities and unknowns in multisensor data fusion(Probabilites et incertitude en fusion de donnees multi-senseurs) ». In : *Revue Scientifique et Technique de la Defense, 1 st Quarter, 1991*, (1991), p. 27-40.

[3] Catriel BEERI et al. « Object fusion in geographic information systems ». In : *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. 2004, p. 816-827.

[4] Mikhail BILENKO et Raymond J MOONEY. « Employing Trainable String Similarity Metrics for Information Integration. » In : *IIWeb*. 2003, p. 67-72.

[5] Isabelle BLOCH et Laurence CHOLVY. « Fusion de données numériques et d'informations symboliques ». In : *TSI. Technique et science informatiques* 22.7-8 (2003).

[6] Bernadette BOUCHON-MEUNIER. « La logique floue et ses applications ». In : (1995).

[7] O COLOT. « Systèmes de perception d'informations incertaines-Application au diagnostic médical ». In : *HDR dissertation, University of Rouen* (2000).

[8] Benjamin COMMANDRE et al. « Manhole Cover Localization in Aerial Images with a Deep Learning Approach ». In : *ISPRS Hannover Workshop : HRIGI 17 – CMRT 17 – ISA 17 – EuroCOW 17*. T. XLII-1/W1. Hannover, Germany, juin 2017, p. 333-338. DOI : 10.5194/isprs-archives-XLII-1-W1-333-2017. URL : https://hal.archives-ouvertes.fr/hal-01556762.

[9] Benoit COSTES. « Matching Old Hydrographic Vector Data from Cassini's Maps ». In : *e-Perimetron* 9.2 (2014), p. 51-65.

[10] Benoit COSTES et Julien PERRET. « A hidden Markov model for matching spatial networks ». In : *Journal of Spatial Information Science* 2019.18 (2019), p. 57-89.

[11] Arthur P DEMPSTER. « Upper and lower probabilities induced by a multivalued mapping ». In : *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 2008, p. 57-72.

[12] Sebastien DESTERCKE. « Independence concepts in evidence theory : some results about epistemic irrelevance and imprecise belief functions ». In : *Workshop on the Theory of Belief Functions*. 2010.

[13] Thomas DEVOGELE, Christine PARENT et Stefano SPACCAPIETRA. « On spatial database integration ». In : *International Journal of Geographical Information Science* 12.4 (1998), p. 335-352.

[14] Didier DUBOIS et Henri PRADE. « La fusion d'informations imprécises ». In : *TS. Traitement du signal* 11.6 (1994), p. 447-458.

[15] Didier DUBOIS et Henri PRADE. « On the unicity of Dempster rule of combination ». In : *International Journal of Intelligent Systems* 1.2 (1986), p. 133-142.

[16] Didler DUBOIS et Henri PRADE. « Representation and combination of uncertainty with belief functions and possibility measures ». In : *Computational intelligence* 4.3 (1988), p. 244-264.

[17] *HydroScience Montpellier*. URL : `http://www.hydrosciences.org/`. (accessed : 26.05.2021).

[18] Radim JIROUŠEK et Jiřina VEJNAROVÁ. « Compositional models and conditional independence in evidence theory ». In : *International Journal of Approximate Reasoning* 52.3 (2011), p. 316-334.

[19] Anne-Laure JOUSSELME et Patrick MAUPIN. « Distances in evidence theory : Comprehensive survey and generalizations ». In : *International Journal of Approximate Reasoning* 53.2 (2012), p. 118-145.

[20] Jung Ok KIM et al. « A new method for matching objects in two different geospatial datasets based on the geographic context ». In : *Computers & Geosciences* 36.9 (2010), p. 1115-1122.

[21] Frank KLAWONN et Erhard SCHWECKE. « On the axiomatic justification of Dempster's rule of combination ». In : *International Journal of Intelligent Systems* 7.5 (1992), p. 469-478.

[22] Alex KRIZHEVSKY, Ilya SUTSKEVER et Geoffrey E HINTON. « Imagenet classification with deep convolutional neural networks ». In : *in Advances in neural information processing systems* (2012), p. 1097-1105.

[23] L LI et MF GOODCHILD. « Automatically and accurately matching objects in geospatial datasets ». In : *Adv. Geo-Spat. Inf. Sci* 10 (2012), p. 71-79.

[24] Linna LI et Michael F GOODCHILD. « An optimisation model for linear feature matching in geographical data conflation ». In : *International Journal of Image and Data Fusion* 2.4 (2011), p. 309-328.

[25] Arnaud MARTIN, Anne-Laure JOUSSELME et Christophe OSSWALD. « Conflict measure for the discounting operation on belief functions ». In : *2008 11th International conference on information fusion*. IEEE. 2008, p. 1-8.

[26] Marie-Hélène MASSON. « Apports de la théorie des possibilités et des fonctions de croyance à l'analyse de données imprécises ». In : *Habilitation à diriger des Recherches soutenue le* 2 (2005).

[27] Sébastien MUSTIÈRE. « Intégration de données géographiques ». Thèse de doct. Université de Grenoble, 2014.

[28] Ana-Maria OLTEANU. « A multi-criteria fusion approach for geographical data matching ». In : *Quality Aspects in Spatial Data Mining* (2009), p. 47.

[29] Ana-Maria OLTEANU. « Fusion de connaissances imparfaites pour l'appariement de données géographiques : proposition d'une approche s' appuyant sur la théorie des fonctions de croyance ». Thèse de doct. Université Paris-Est, 2008.

[30] Ana-Maria OLTEANU-RAIMOND, Sebastien MUSTIERE et Anne RUAS. « Knowledge formalization for vector data matching using belief theory ». In : *Journal of Spatial Information Science* 2015.10 (2015), p. 21-46.

[31] Zdzisław PAWLAK. « Rough sets ». In : *International journal of computer & information sciences* 11.5 (1982), p. 341-356.

[32] Barbara ROSEN et Alan SAALFELD. « Match criteria for automatic alignment ». In : *Proceedings of 7th international symposium on computer-assisted cartography (Auto-Carto 7)*. 1985, p. 1-20.

[33] Cyril ROYÈRE. « Contribution à la résolution du conflit dans le cadre de la théorie de l'évidence : Application à la perception et à la localisation des véhicules intelligents ». Thèse de doct. Compiègne, 2002.

[34] A SAALFELD. « Conflation : Automated Map Compilation. Computer Vision Laboratory ». In : *Center for Automation Research, University of Maryland* (1993).

[35] Alan SAALFELD. « Conflation automated map compilation ». In : *International Journal of Geographical Information System* 2.3 (1988), p. 217-228.

[36] Ashok SAMAL, Sharad SETH et Kevin CUETO 1. « A feature-based approach to conflation of geospatial sources ». In : *International Journal of Geographical Information Science* 18.5 (2004), p. 459-489.

[37] Ahmed SAMET, Eric LEFEVRE et Sadok Ben YAHIA. « Reliability estimation with extrinsic and intrinsic measure in belief function theory ». In : *2013 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO)*. IEEE. 2013, p. 1-6.

[38] Ahmed SAMET et al. « Generic discounting evaluation approach for urban image classification ». In : *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*. Springer. 2013, p. 79-90.

[39] G SAPORTA. « Probabilités, Analyse des données et statistique : Technip ». In : *Paris, France* (1990).

[40] G SHAFER. « A Mathematical Theory of Evidence. Princeton University Press, Princeton, N. J., 1976 ». In : *ShaferA Mathematical Theory of Evidence1976* ().

[41] Philippe SMETS. « Analyzing the combination of conflicting belief functions ». In : *Information fusion* 8.4 (2007), p. 387-412.

[42] Philippe SMETS. « Belief functions : the disjunctive rule of combination and the generalized Bayesian theorem ». In : *International Journal of approximate reasoning* 9.1 (1993), p. 1-35.

[43] Philippe SMETS. « Decision making in the TBM : the necessity of the pignistic transformation ». In : *International journal of approximate reasoning* 38.2 (2005), p. 133-147.

[44] Philippe SMETS. « Imperfect information : Imprecision and uncertainty ». In : *Uncertainty management in information systems*. Springer, 1997, p. 225-254.

[45] Philippe SMETS. « Non-standard logics for automated reasoning ». In : (1988).

[46] Philippe SMETS et Robert KENNES. « The transferable belief model ». In : *Artificial intelligence* 66.2 (1994), p. 191-234.

[47] Wenbo SONG et al. « Relaxation-based point feature matching for vector map conflation ». In : *Transactions in GIS* 15.1 (2011), p. 43-60.

[48] Gábor J SZÉKELY. *Paradoxes in probability theory and mathematical statistics*. Rapp. tech. 1986.

[49] Xiaohua TONG, Wenzhong SHI et Susu DENG. « A probability-based multi-measure feature matching method in map conflation ». In : *International Journal of Remote Sensing* 30.20 (2009), p. 5453-5472.

[50] Steffen VOLZ. « An iterative approach for matching multiple representations of street data ». In : *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 36.Part 2/W40 (2006), p. 101-110.

[51] Volker WALTER et Dieter FRITSCH. « Matching spatial data sets : a statistical approach ». In : *International Journal of geographical information science* 13.5 (1999), p. 445-473.

[52] Yanxia WANG et al. « A back-propagation neural network-based approach for multi-represented feature matching in update propagation ». In : *Transactions in GIS* 19.6 (2015), p. 964-993.

[53] Jianhua WU et al. « A matching algorithm based on Voronoi diagram for multi-scale polygonal residential areas ». In : *IEEE Access* 6 (2018), p. 4904-4915.

[54] Emerson MA XAVIER, Francisco J ARIZA-LÓPEZ et Manuel A URENA-CAMARA. « A survey of measures and methods for matching geospatial vector datasets ». In : *ACM Computing Surveys (CSUR)* 49.2 (2016), p. 1-34.

[55] Boutheina Ben YAGHLANE, Philippe SMETS et Khaled MELLOULI. « Belief function independence : I. the marginal case ». In : *International Journal of Approximate Reasoning* 29.1 (2002), p. 47-70.

[56] Deng YONG et al. « Combining belief functions based on distance of evidence ». In : *Decision support systems* 38.3 (2004), p. 489-493.

[57] Shuxin YUAN et Chuang TAO. « Development of conflation components ». In : *Proceedings of Geoinformatics* 99 (1999), p. 1-13.

[58] Lotfi A ZADEH. « A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination ». In : *AI magazine* 7.2 (1986), p. 85-85.

[59] Lotfi A ZADEH. « Information and control ». In : *Fuzzy sets* 8.3 (1965), p. 338-353.

[60] Lotfi A ZADEH. « Review of a mathematical theory of evidence ». In : *AI magazine* 5.3 (1984), p. 81-81.

[61] Lotfi Asker ZADEH. « Fuzzy sets as a basis for a theory of possibility ». In : *Fuzzy sets and systems* 1.1 (1978), p. 3-28.

[62] Jianchen ZHANG, Yanhui WANG et Wenji ZHAO. « An improved hybrid method for enhanced road feature selection in map generalization ». In : *ISPRS International Journal of Geo-Information* 6.7 (2017), p. 196.

[63] Meng ZHANG, Wei SHI et Liqiu MENG. « A generic matching algorithm for line networks of different resolutions ». In : *Workshop of ICA commission on generalization and multiple representation computering faculty of a Coruña University-Campus de Elviña, Spain*. T. 9. Citeseer. 2005, p. 101-110.

Annexes

Annexe A

Définitions

A.1 Les formes d'imperfection d'information

En général, dans le domaine de l'intelligence artificielle, la notion d'imperfection de l'information fait appel à trois concepts : l'imprécision, l'incertitude et l'incomplétude (figure A.1) [6, 7, 26, 44].

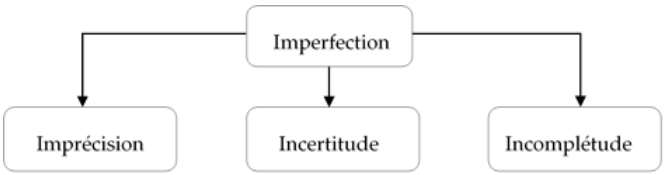


FIGURE A.1 – Grands types d'imperfection

Le premier concept est l'imprécision qui concerne une difficulté dans l'énoncé des connaissances soit parce que :

- Des expressions du langage naturel sont utilisées pour décrire certaines caractéristiques. Par exemple, Les différents stades de la vie d'un Homme (Bébé, Enfant, Adolescent, adulte, personne âgée).
- Des connaissances numériques sont mal connues, dans notre cas, par exemple, diamètre d'une conduite entre la valeur 150 et 200.

Le deuxième concept est l'incertitude qui fait référence à la véracité de l'information : l'information peut être complète et précise mais fausse. Dans notre cas, l'incertitude peut por-

ter par exemple sur l'existence d'un regard ou d'une conduite.

Le troisième concept est l'incomplétude qui correspond à l'absence d'information, dans notre cas, par exemple, le champ type de matériel n'est pas renseigné pour une conduite.

L'imprécision et l'incomplétude selon [16] font référence au contenu de l'information, par contre l'incertitude fait référence non pas au contenu de l'information mais à sa qualité.

Comme nous l'avons vu, les connaissances sont imparfaites. Par conséquent, dans n'importe quel système expert il est essentiel d'utiliser des formalismes qui nous permettent de modéliser les différentes formes d'imperfection de l'information. Il faut avoir également des mécanismes de fusion des informations dans ces formalismes ; cette étape a pour rôle de synthétiser et résumer toutes les connaissances.

A.2 QGIS

QGIS est une application SIG [1] gratuite et multiplateforme qui permet de créer, modifier, visualiser et analyser des données géographiques. Supportant un grand nombre de formats de données (Raster, Vector, Tiff, etc.), QGIS fait partie des meilleures solutions pour la gestion des objets géographiques.

A.3 Les types de données SIG

Dans le monde du SIG, deux types de données sont les plus utilisées : les données Vecteur et Raster.

A.3.1 Les données Raster

Pour représenter le monde réel, les données Raster utilisent une matrice de pixels, dans laquelle chaque pixel représente un morceau d'information. Ce type est habituellement utilisé pour représenter des informations continues qui ne sont pas facilement séparables à l'aide de la représentation vectorielle (figure A.2).

1. *Système d'Information Géographique*

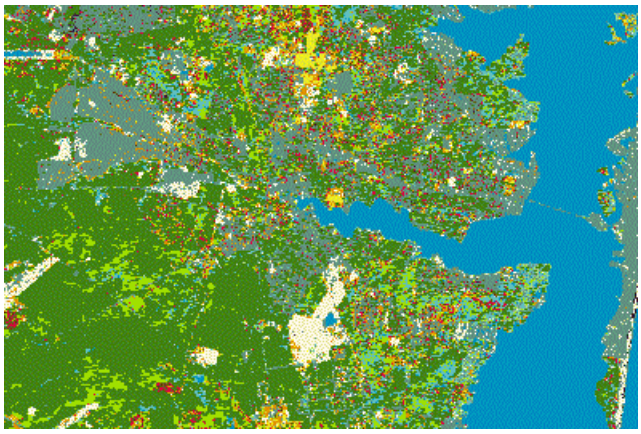


FIGURE A.2 – Exemple d'une couche Raster

A.3.2 Les données Vecteur

Les données vecteurs sont utilisées pour représenter les caractéristiques du monde réel dans un SIG. Une entité vecteur peut avoir une géométrie de type :

- **point** : lorsqu'une géométrie est composée d'un seul sommet.
- **polyligne** : quand une géométrie est composée de deux ou plusieurs sommets reliés par une ligne.
- **polygone** : lorsqu'une géométrie est construite avec au moins quatre sommets formant une surface fermée. Le premier et le dernier sommet sont toujours au même endroit.

Dans notre exemple de la Figure A.3, la représentation est la suivante : les Points pour les ouvrages, le Polylines pour les routes et le Polygones pour les maisons.

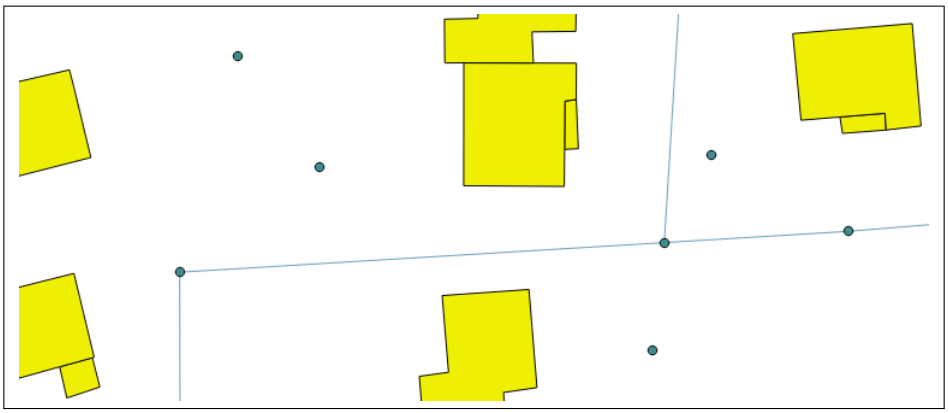
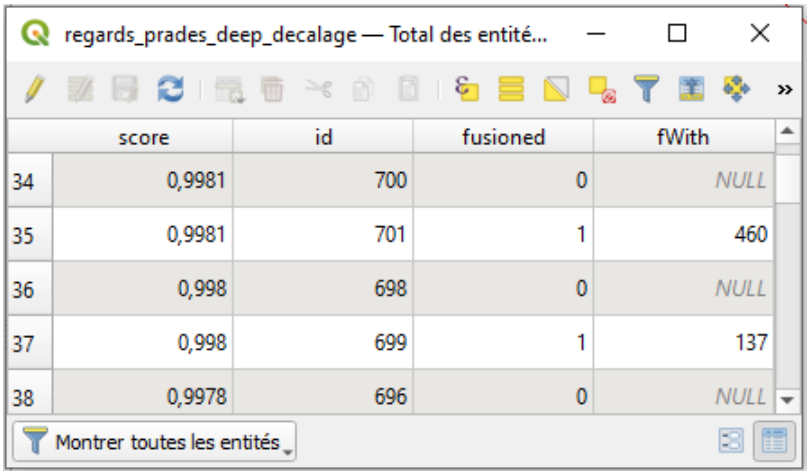


FIGURE A.3 – Exemple de données Vecteur

Chaque entité vectorielle possède des données attributaires qui la décrivent.Par exemple, l'ID, score etc. (Figure A.4)



	score	id	fusioned	fWith
34	0,9981	700	0	NULL
35	0,9981	701	1	460
36	0,998	698	0	NULL
37	0,998	699	1	137
38	0,9978	696	0	NULL

FIGURE A.4 – Exemple d'une table attributaire

A.4 Le stockage de données SIG

Pour le stockage des données, QGIS offre principalement deux possibilités :

- La première option est de stocker les données dans des fichiers situés sur le disque de l'ordinateur. Il y a un certain nombre de formats de fichiers différents pour les données SIG, mais le plus courant est probablement le "shapefile", qui est en fait constitué d'au moins trois fichiers distincts qui fonctionnent ensemble pour stocker vos données vectorielles numériques :
  - **.shp** : La géométrie des entités vectorielles est stockée dans ce fichier
  - **.dbf** : Les attributs des entités vectorielles sont stockés dans ce fichier
  - **.shx** : Ce fichier est un index qui aide les applications SIG à rapidement identifier les entités.
  - **.prj optionnel mais recommande** : contient la projection liée aux entités afin qu'elles soient géo-référencées.
- la seconde option utilise des bases de données (SGBD) pour stocker les données plus efficacement. L'application QGIS offre deux types de bases de données sur lesquelles nous pouvons stocker nos données, PostGIS et SpatialLite. Cependant, cela ne veut pas dire que nous ne pouvons pas utiliser d'autres SGBD.

A.5 GDAL

GDAL et une bibliothèque open source qui permet de lire et de traiter un grand nombre de formats d'images géographiques à partir de plusieurs langages de programmation tels que C, C++, JAVA, Python, ou le langage statistique R, etc. Un sous-ensemble de cette bibliothèque est la bibliothèque ORG, qui donne accès aux formats de données vectorielles les plus courants. Cette bibliothèque est l'un des piliers des systèmes d'information géographique open source, puisqu'elle assure la compatibilité avec de nombreux systèmes commerciaux basés sur des formats propriétaires ainsi qu'avec les normes de l'Open Geospatial Consortium.

Annexe B

Les formalismes de gestion des données imparfaites

B.1 La théorie des probabilités

Historiquement, le formalisme le plus utilisé est la théorie des probabilités. Malheureusement, ce formalisme ne permet pas de modéliser facilement les différentes formes d'imperfections mentionnées précédemment.

Un des premiers points soulevés concerne la représentation de l'ignorance liée au principe de raison insuffisante, qui revient à accorder la même probabilité à chaque événement en l'absence d'information, parce que c'est la probabilité la moins fausse. A partir de cette modélisation, on peut avoir deux notions totalement différentes, qui conduisent pourtant à la même probabilité. Par exemple [1], dans la compétition de tir à la carabine à 10 mètres des Jeux Olympiques de Londres 2012. Parmi les compétiteurs de cette épreuve, nous nous intéressons plus particulièrement à 3 compétiteurs : Niccolò Campriani (tireur italien, abréviation NC), Alin George Moldoveanu (tireur roumain, abréviation AGM) et Pierre Edmond Piasecki (tireur français, abréviation PEP). Quand on interroge un spécialiste qui maîtrise parfaitement cette discipline, il affirme que ces trois concurrents, à palmarès équivalent, ont les mêmes chances de gagner la médaille d'or. Cela conduit à la probabilité suivante $P(NC) = P(AGM) = P(PEP) = \frac{1}{3}$, nous interrogeons maintenant quelqu'un qui n'a aucune connaissance de cette discipline. Sans aucune référence et sans aucune expérience dans ce domaine, la probabilité la moins erronée est $P(NC) = P(AGM) = P(PEP) = \frac{1}{3}$. Le second problème

concerne l'axiome de l'additivité : il est impossible de donner une probabilité à un événement sans influencer son complémentaire [39, 48]. Par exemple, l'attribution d'une probabilité x à un événement A associe automatiquement une probabilité $1 - x$ à l'événement "non A".

B.2 La théorie des ensembles approximatifs

La théorie des ensembles approximatifs, élaborée par [31] au début des années 1980, est axée sur la notion d'ensembles approximatifs. Un ensemble approximatif f est alors défini par une fonction d'appartenance qui attribue à chaque élément $x \subseteq X$ une certaine valeur dans l'ensemble E comme suit :

$$f : X \rightarrow E, avecE = \{T, M, F\} \tag{B.1}$$

où T, M et F représentent respectivement "appartient peut-être", "appartient" et "n'appartient pas".

Les concepts sont représentés en fonction de leurs approximations inférieures et supérieures. De ce fait, l'approximation supérieure d'une classe X est l'ensemble des objets qui ne font certainement pas partie de X. L'approximation inférieure d'une classe X est l'ensemble des objets qui en font partie.

La théorie des ensembles approximatifs permet de traiter l'imprécision et l'incertitude propres aux situations de décision. À la différence de l'approche des probabilités conditionnelles, elle ne nécessite pas de connaissance a priori des données.

B.3 La théorie des ensembles flous

La théorie des (sous-)ensembles flous est une généralisation de la théorie des ensembles classiques (ou ensembles nets). Cette théorie a été introduite par Zadeh [59], Ce dernier a présenté la notion de sous-ensemble flou pour résoudre le problème posé par les connaissances exprimées symboliquement, imprécises ou vagues.

Un sous-ensemble A d'un référentiel Ω est classiquement défini par les éléments qui le composent, donc un élément x appartient ou n'appartient pas à A, c'est-à-dire la proposition

logique associée "l'élément x appartient à l'ensemble" A est soit fausse, soit vraie. Ce concept d'ensemble classique a été adapté à celui d'ensemble flou par la notion d'appartenance partielle ou de vérité partielle. Un élément x est alors membre d'un sous-ensemble flou A avec un degré de d'appartenance $\mu_A(x)$. La fonction d'appartenance $\mu_A : \Omega \rightarrow [0, 1]$ attribue à chaque élément du référentiel Ω un degré d'appartenance.

Cette théorie des sous-ensembles flous présente l'avantage de pouvoir modéliser les opérateurs du type ET, OU et également de les combiner mais elle ne permet pas la modélisation de l'incertitude. Cependant, comme le précise Bouchon-Meunier [6], l'imprécision et l'incertitude sont étroitement liées. Par conséquent, [61] a introduit la théorie des possibilités.

B.4 La théorie des possibilités

Zadeh [61] a d'abord introduit la théorie des possibilités en 1978 comme une extension de sa théorie des ensembles flous et la logique floue. Dubois et BLOCH [16, 5] ont ensuite contribué à son développement. Cette théorie assure la modélisation des incertitudes sur les événements, sans nécessiter de connaissances a priori. Elle permet de dire dans quelle mesure un événement est possible et dans quelle mesure il est certain. Ces deux affirmations sont formalisées par deux mesures : la mesure de possibilité et la mesure de nécessité. Ces deux mesures sont des cas particuliers de plausibilité et de crédibilité.

B.4.1 La mesure de possibilité

Pour un ensemble de référence Ω contenant un certain nombre d'événements, on affecte à chaque événement, autrement dit à chaque élément de l'ensemble $P(\Omega)$ des sous-ensembles ordinaires de Ω, une mesure de la possibilité de cet événement au moyen d'une valeur comprise entre 0 et 1.

Définition : une mesure de possibilité [6], notée Π, est une fonction définie sur l'ensemble $P(\Omega)$, prenant des valeurs dans l'intervalle $[0, 1]$, telle que :

$$\begin{cases} \Pi(\Phi) = 0, \Pi(\Omega) = 0 \\ \forall A_i \in P(\Omega), i = 1, 2, \ldots, n \ \Pi(\cup_{i=1,2,\ldots,n}) = \sup_{i=1,2,\ldots,n} \Pi(A_i) \end{cases} \tag{B.2}$$

où sup représente la plus grande valeur dans le cas fini.

B.4.2 La mesure de nécessité

La mesure de nécessité désigne le niveau de certitude de la réalisation d'un événement, en affectant un coefficient entre 0 et 1 à toute partie de Ω. Les propriétés de la mesure de nécessité, différentes de celles de la mesure de possibilité, sont les suivantes :

Définition : une mesure de nécessité, notée N, est une fonction définie sur l'ensemble $P(\Omega)$, prenant des valeurs dans l'intervalle $[0, 1]$, telle que :

$$\begin{cases} \Pi(\Phi) = 0, \Pi(\Omega) = 0 \\ \forall A_i \in P(\Omega), i = 1, 2, \ldots, n \; N(\cap_{i=1,2,\ldots}) = \inf_{i=1,2,\ldots,n} \Pi(A_i) \end{cases} \tag{B.3}$$

Où $\inf$ représente la plus petite valeur parmi toutes les valeurs, dans le cas fini.

Ainsi, une mesure de nécessité combine l'intersection de deux événements avec la plus petite valeur des coefficients assignés à chaque événement.

La mesure de nécessité est associée à la mesure de possibilité par la formule :

$$\Pi(A) = 1 - N(\overline{A}) \tag{B.4}$$

La théorie des possibilités offre un grand nombre de types d'opérateurs de fusion [14]. Ainsi, on peut citer notamment l'opérateur minimal, qui effectue une combinaison conjonctive lorsque les sources d'information sont accordées, l'opérateur maximal, qui effectue une combinaison disjonctive lorsque les sources sont désaccordées, et l'opérateur adaptatif, qui réalise une combinaison adaptative en fonction du niveau de compatibilité des deux sources d'information.

Une différence importante entre la théorie des probabilités et la théorie des possibilités concerne un événement et son complément. La probabilité d'un événement dépend entièrement de son complément, tandis que la possibilité et la nécessité d'un événement et de son complément sont faiblement liées.

La théorie des possibilités modélise la prédilection pour une hypothèse, alors que la théorie des fonctions de croyance formalise un degré de vérité plus ou moins certain.

B.5 La théorie des fonctions de croyance

Un autre cadre théorique très intéressant est certainement la théorie des fonctions de croyance. L'un des principaux ambassadeurs de ce formalisme est sans doute Ph. Smets qui a, via le modèle des croyances transférables [45, 46], continué à faire évoluer la théorie développée par A. Dempster [11] et G. Shafer [40]. Cette théorie permet de représenter la connaissance d'une manière relativement naturelle et de modéliser les diverses formes d'imperfection de l'information. De plus, les bases de cette théorie sont très simples à comprendre, ce qui représente un avantage important. Enfin, elle unifie toutes les théories de l'incertitude. C'est pour ces raisons que nous avons essentiellement fait appel à ce formalisme dans ce projet.

Nous avons présenté dans le chapitre 2, certains éléments mathématiques associés à ce formalisme qui a été utilisé pour traiter de nombreux problèmes liés au traitement des informations imparfaites.

Fusion de données spatiales : Gestion des données imparfaites en utilisant les fonctions de croyance

Résumé

De nos jours, les données concernant les réseaux d'assainissement qui couvrent le même territoire sont disponibles à partir de plusieurs sources. Cette diversité de sources a donné lieu à plusieurs travaux axés sur le thème de la visualisation, l'analyse et la fusion de bases de données spatiales à des fins diverses telles que l'amélioration de la qualité ou la fourniture de nouveaux services. Pour rendre plus explicites les relations entre les différents objets des bases de données, nous utilisons le processus d'appariement des données géographiques qui consiste à mettre en correspondance les objets homologues représentant la même réalité. La complexité du processus d'appariement implique que les approches existantes varient en fonction des besoins que la correspondance doit satisfaire, et varient en fonction des types de données à mettre en correspondance (points, lignes ou surfaces) et du degré de détail. Nous avons trouvé que la plupart des approches sont basées sur la géométrie et leurs relations et que très peu prennent en charge les connaissances sémantiques. Pour tenir compte de tous les critères géométriques et sémantiques et leurs relations, nous utilisons la théorie des fonctions de croyance, dans le but de formaliser la connaissance et de modéliser les imperfections des données rencontrées lors de la fusion des données des réseaux d'assainissement multi-source.

Mots-clés : théorie des fonctions de croyance, appariement, fusion, connaissances, imperfection.

Spatial data fusion: Managing imperfect data using belief functions

Abstract

Nowadays, data concerning sewerage networks covering the same territory are available from various sources. This diversity of sources has given rise to several works focusing on the visualisation, analysis and fusion of spatial databases for various purposes such as quality enhancement or the provision of new services. To make the relationships between the different objects in the databases more explicit, we use the process of geographic data matching, which consists in matching homologous objects representing the same reality. The complexity of the matching process means that existing approaches vary according to the needs that the matching must satisfy, and vary according to the kinds of data to be matched (points, lines or surfaces) and the degree of detail. We have found that most approaches are based on geometry and its relationships and very few supports semantic knowledge. In order to take into account all the geometric and semantic criteria and their relations, we use the belief functions theory, with the aim of formalising the knowledge and modelling the data imperfections encountered when fusing data from multi-source sewerage networks.

Keywords: belief functions theory, matching, fusion, knowledge, imperfection.

Master Systèmes Intelligents & Réseaux
Département d'Informatique
Faculté des Sciences et Techniques de Fès
A.U. 2020 - 2021