Estimation de paramètres clés liés à la gestion d'un réseau de distribution d'eau potable : Méthode d'inférence sur les noeuds d'un graphe

Christophe Dumora

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR

DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET

INFORMATIQUE

Mathématiques Appliquées et Calcul Scientifique

Par **Christophe DUMORA**

Estimation de paramètres clés liés à la gestion d'un réseau de distribution d'eau potable : Méthode d'inférence sur les nœuds d'un graphe

Soutenue le : 15 décembre 2020

Membres du jury :

Mme.	VIALANEIX Nathalie	Directrice de recherche	INRAE Toulouse	Présidente
Mme.	PEYRARD Nathalie	Directrice de recherche	INRA Toulouse	Rapportrice
M.	PUDLO Pierre	Professeur	Université Aix-Marseille	Rapporteur
M.	BIGOT Jérémie	Professeur	Université de Bordeaux	Directeur
M.	AUBER David	Professeur	Université de Bordeaux	Co-directeur
M.	COUALLIER Vincent	Maître de conférence	Université de Bordeaux	Encadrant
	Membres invités :			
M.	LECLERC Cyril	Chef de projet	Le LyRE SUEZ Eau France	Invité
M.	PILLER Olivier	Chargé de recherche	INRAE Bordeaux	Invité

Estimation de paramètres clés liés à la gestion d'un réseau de distribution d'eau potable : Méthode d'inférence sur les nœuds d'un graphe

Résumé :

L'essor des données générées par les capteurs et par les outils opérationnels autour de la gestion des réseaux d'alimentation en eau potable (AEP) rendent ces systèmes de plus en plus complexes et de façon générale les événements plus difficiles à appréhender. L'historique de données lié à la qualité de l'eau distribuée croisé avec la connaissance du patrimoine réseau, des données contextuelles et des paramètres temporels amène à étudier un système complexe de par sa volumétrie et l'existence d'interactions entre ces différentes données de natures diverses pouvant varier dans le temps et l'espace. L'utilisation de graphes mathématiques permet de regrouper toute cette diversité et fournit une représentation complète des réseaux AEP ainsi que les évènements pouvant y survenir ou influer sur leur bon fonctionnement. La théorie des graphes associées à ces graphes mathématiques permet une analyse structurelle et spectrale des réseaux ainsi constitués afin de répondre à des problématiques métiers concrètes et d'améliorer des processus internes existants. Ces graphes sont ensuite utilisés pour répondre au problème d'inférence sur les noeuds d'un très grand graphe à partir de l'observation partielle de quelques données sur un faible nombre de noeuds. Une approche par algorithme d'optimisation sur les graphes est utilisée pour construire une variable numérique de débit en tout noeuds du graphe (et donc en tout point du réseau physique) à l'aide d'algorithme de flots et des données issues des débitmètres réseau. Ensuite une approche de prédiction par noyau reposant sur un estimateur pénalisé de type Ridge, qui soulève des problèmes d'analyse spectrale de grande matrice creuse, permet l'inférence d'un signal observé sur un certain nombre de noeuds en tout point d'un réseau AEP.

Mots-clés : Inférence statistique, Théorie des graphes, Réseau de distribution d'eau potable, Régression Ridge à noyau, Algorithme de flots

---

Estimation of key parameters related to the management of a drinking water distribution network : Inference method on the nodes of a graph

Abstract : The rise of data generated by sensors and operational tools around water distribution network (WDN) management make these systems more and more complex and in general the events more difficult to predict. The history of data related to the quality of distributed water crossed with the knowledge of network assets, contextual data and temporal parameters lead to study a complex system due to its volume and the existence of interactions between these various type of data which may vary in time and space. This big variety of data is grouped by the use of mathematical graph and allow to represent WDN as a whole and all the events that may arise therein or influence their proper functioning. The graph theory associated with these mathematical graphs allow a structural and spectral analysis of WDN to answer to specific needs and enhance existing process. These graphs are then used to answer the probleme of inference on the nodes of large graph from the observation of data on a small number of nodes. An approach by optminisation algorithm is used to construct a variable of flow on every nodes of a graph (therefore at any point of a physical network) using flow algorithm and data measured in real time by flowmeters. Then, a kernel prediction approach based on a Ridge estimator, which raises spectral analysis problems of a large sparse matrix, allow the inference of a signal measured on specific nodes of a graph at any point of a WDN.

Keywords : Statistical inference, Graphe theory, Water distribution network, Kernel Ridge regression, Flow algorithm

---

Remerciements

Je tiens tout d'abord à remercier Mme Nathalie Peyrard et Mr Pierre Pudlo qui m'ont fait l'honneur de rapporter ce manuscrit de thèse. Je remercie aussi Mme Nathalie Vialaneix ainsi que Mr Olivier Piller d'avoir accepté de faire partie de mon jury de thèse.

Je tiens à remercier l'ensemble des personnes que j'ai pu rencontrer et avec qui j'ai pu échanger durant toutes ces années au LyRE. Je garde un remerciement tout particulier pour Cyril Leclerc sans qui je n'aurais jamais eu l'opportunité d'initier ces travaux de thèse. Merci de m'avoir fait confiance pour cette longue aventure qu'a été la thèse, j'ai beaucoup appris à tes côtés professionnellement et humainement. Un grand merci à Thierry qui a été mon premier collègue de bureau pendant que j'étais en stage toujours à l'écoute et plein de bons conseils. Merci à Mélodie et Damien pour la confiance et le soutien que vous avez toujours su m'accorder. Merci aussi à Karim et Julien qui ont été des exemples à suivre durant ma thèse au LyRE. Je ne pourrais citer tout le monde mais merci à vous tous d'avoir contribué de près ou de loin à la réussite de ces travaux.

Merci aussi à mes encadrants qui m'ont accompagné tout au long de cette thèse. Merci à Jérémie Bigot mon directeur de thèse qui a toujours su m'aiguiller et me rassurer sur l'avancée de mes travaux. Merci pour ces discussions dans ton bureau à imaginer des choses toujours plus folles avec ce sujet de recherche. Merci à David mon second directeur de thèse qui a su m'ouvrir les portes de la recherche en informatique. Merci pour ton temps et ces sessions de code dans ton bureau. Un remerciement tout particulier à Vincent Couallier mon co-encadrant de thèse. Je t'ai connu en tant que professeur et mentor à la fac et c'est en tant que collègue et ami que j'ai eu la chance de te côtoyer. Tu as toujours été d'un soutien sans faille et ce même lorsque j'ai pu avoir des doutes sur l'avancée de ma thèse. Merci pour ton aide durant de cette longue et fastidieuse ligne qu'a été la rédaction de manuscrit. A tous les trois merci pour votre bienveillance, votre écoute et votre soutien.

Mes remerciements vont aussi à ma famille et mes amis qui, avec cette question récurrente, "Quand est-ce que tu la finis cette thèse ?", bien qu'angoissante en période fréquente de doutes, m'a permis de jamais dévier de mon objectif final. Merci à Manon, Sarah, Vanessa, Nicolas, Yannick, Steven, Benjamin, Tony, Lucas, Jessica, Elodie, Corine, Stephane, Xavier, Jean-Daniel, Chantal, Lucile, Clémence, Kévin, Matias pour leur affection maintes fois renouvelée.

Merci à mes parents, Isabelle et Alain, qui m'ont toujours aidé, soutenu et supporté dans tout ce que j'ai entrepris, votre présence et vos encouragements sont pour moi les piliers fondateurs de ce que je suis et de ce que je fais.

Enfin, je remercie ma compagne et meilleure amie de toujours Claire pour son soutien quotidien indéfectible durant cette aventure qui n'a pas toujours été facile à vivre. Sans toi cette thèse n'aurait pas eu le même goût.

Table des matières

Table des figures

Liste des tableaux

Notations

Nous présentons ci-dessous les différentes notations mathématiques générales utilisées à travers le manuscrit et leurs significations. Des notations plus spécifiques sont introduites là où elles sont nécessaires.

G	Graphe mathématique
V	Ensemble de sommets d'un graphe, ici de taille finie
N_v	Nombre de sommets d'un graphe de taille finie : $N = \|V\|$
E	Ensemble d'arêtes d'un graphe non orienté, de taille finie : $E \subseteq V \times V$
N_e	Nombre d'arêtes d'un graphe de taille finie : $M = \|E\|$
$d(v)$	Degrés du sommet $V \in E$
$\Gamma_{(v)}$	Ensemble voisin d'un sommet $v \in E$: sommets partageant une arête avec v
$V(S)$	Ensemble des sommets d'un sous-graphe S donné
$E(S)$	Ensemble des arêtes d'un sous-graphe S donné
MOD	Modèle orienté donné
$PRESS$	Predicted Error sum of squares
GCV	Validation croisée généralisée

Glossaire

AEP	Réseau d'Alimentation en Eau Potable
BM	Bordeaux Métropole
VLAR	Volumes Livrés au Réseau
TLRV	Compteur Télé-relevé
CC	Composante connexe
LOOCV	Leave-one-out cross validation

Introduction

L'eau est une ressource naturelle, un bien commun qui en lui-même n'a pas de prix. Ce sont les efforts quotidiens mis en œuvre pour acheminer une eau potable de qualité aux robinets des consommateurs qui en déterminent le coût. Être capable de fournir une eau potable conforme aux règles sanitaires est une préoccupation permanente pour une société comme Suez. L'enjeu d'aujourd'hui est bien évidemment de distribuer une eau conforme aux exigences sanitaires mais possédant également une qualité gustative irréprochable.

SUEZ a instrumenté ses réseaux depuis maintenant plusieurs années. Les capteurs positionnés sur les usines de production, le réseau de distribution ou au plus près du consommateur permettent de suivre et de contrôler des paramètres clés liés à la qualité de l'eau (chlore, pH ou encore température) ou à la demande en eau (débit, volume consommé) et ceci sur un pas de temps très fin. Cet historique de données lié à la qualité de l'eau distribuée croisé avec la connaissance du patrimoine réseau, des données contextuelles et des paramètres temporels amène à étudier un système complexe de par sa volumétrie et l'existence d'interactions entre ces différentes données de natures diverses pouvant varier dans le temps et l'espace.

L'essor des données générées par les capteurs et par les outils opérationnels rendent les systèmes de plus en plus complexes et de façon générale les événements plus difficiles à appréhender. Il est ainsi compliqué de connaitre et anticiper les variations de flux hydrauliques au cours du temps en tout point du réseau.

Nous présentons dans le premier chapitre, le contexte de villes intelligentes. Après une brève présentation du petit cycle de l'eau, représentant le cheminement de l'eau depuis son extraction jusqu'au robinet du consommateur, nous discutons de la notion de réseaux d'eau intelligents. Le réseau d'alimentation en eau potable (AEP) étudié durant cette thèse est celui de la Métropole Bordelaise. Nous détaillons d'abord les données topologiques du réseau représentant le patrimoine enterré de la Métropole Bordelaise. Ensuite nous présentons l'ensemble des données métrologiques utilisées concernant le suivi de paramètres clés liés à la gestion des réseaux AEP.

L'aspect novateur du travail de recherche tient en l'approche méthodologique choisie : représenter l'ensemble des données générées autour de la gestion d'un réseau AEP sous la forme d'un réseau de graphe. Dans le chapitre 2 nous présentons les notations et concepts de base de la théorie des graphes, ainsi que la méthodologie permettant de prendre en compte les données topologiques et métrologiques issues des différents outils métiers afin de représenter un réseau AEP sous la forme d'un réseau de graphe. Regrouper toute la diversité des données permettant une représentation complète des réseaux AEP ainsi que les évènements pouvant avoir lieux sur ceux-ci ou influer sur leur bon fonctionnement. Nous présentons ensuite une analyse des caractéristiques structurelles et spectrales des

réseaux ainsi constitués. Nous proposons de nouveaux outils pour la conception et la gestion patrimoniale tirant avantage de la structure des données et de la théorie des graphes.

Dans le chapitre 3 nous détaillons un algorithme de flots pour la reconstruction hydraulique d'un réseau d'alimentation en eau potable. Après une introduction de quelques définitions et notations liées aux réseaux de flots nous nous intéressons à la méthode de Ford & Fulkerson permettant de faire circuler un flot dans un réseau de flot. Ensuite nous présentons les spécificités liées à l'adaptation d'une telle méthode à la reconstruction des débits en tout point d'un réseau AEP.

Le dernier chapitre traite de la reconstruction d'un signal partiellement observé sur un graphe par une méthode statistique de régression à noyau. Nous présentons dans un premier temps les notations ainsi que les méthodes nous permettant de définir un noyau à partir du graphe représentant le réseau de distribution d'eau potable. Nous abordons le problème d'inférence sur les nœuds d'un très grand graphe à l'aide d'une approche de prédiction par noyau reposant sur un estimateur pénalisé de type Ridge, soulevant des problèmes d'analyse spectrale d'une très grande matrice creuse. Enfin, nous présentons quelques extensions aux signaux temporels ainsi que les résultats obtenus.

Cette thèse, financée par Bordeaux Métropole, a été réalisée au LyRE, centre de Recherche et Développement de SUEZ et co-encadrée par l'Institut de Mathématiques de Bordeaux, sous la direction de Jérémie Bigot et Vincent Couallier et le Laboratoire Bordelais de Recherche en Informatique sous la direction de David Auber.

Chapitre 1

Contexte

1.1 Villes intelligentes, réseaux de distribution d'eau

L'organisation mondiale de la santé rapporte que la population urbaine représente 54% de la population totale, contre 34% en 1960, et continue de grandir. Le taux d'urbanisation devrait se situer un peu au-dessus de 60% en 2030 [Nations, 2007]. La planète compte aujourd'hui 3.3 milliards de citadins, soit quatre fois et demi plus qu'en 1950. En 2030, l'effectif de la population urbaine devrait atteindre 5 milliards ; il y aura alors autant de citadins dans le monde qu'il n'y avait d'habitants sur Terre en 1987 [Véron, 2006].

La plupart des ressources sont aujourd'hui consommées dans des villes du monde entier, la prise de conscience de cette tendance pousse la recherche de nouvelles façons de répondre aux demandes croissantes. Cette situation incite les villes à trouver des moyens nouveaux et plus intelligents afin de gérer les nouveaux défis, de comprendre et maîtriser les interactions complexes entre les systèmes technologiques et les systèmes de prestations de services et leurs impacts sur la durabilité de l'environnement ([Caragliu et al., 2011], [Mori and Christodoulou, 2012])

C'est dans ce contexte que le concept de « ville intelligente » a été introduit. Depuis ces deux dernières décennies le concept de « ville intelligente » est de plus en plus populaire à la fois dans la littérature scientifique et au sein des politiques internationales. Néanmoins il n'existe pas de définition consensuelle de ce terme, [Albino et al., 2015] en ont listé 23 distinctes. Cette variété s'explique notamment en raison du fait que la « ville intelligente », de par la diversité des domaines qu'elle touche, constitue un objet de recherches multidisciplinaires [Angelidou, 2015]. Néanmoins le sens originel définirait une ville intelligente comme un ville dans laquelle les technologies de l'information et de la communication jouent un rôle essentiel dans l'amélioration de la qualité de la vie et l'atteinte de l'excellence économique [Mahizhnan, 1999].

Les capteurs, les données, les analyses et leurs connectivités offrent la possibilité de garantir le fonctionnement de chacun des services fournis dans les villes en incrustant des dispositifs numériques dans les infrastructures urbaines. Par exemple un meilleur transport peut aider à fournir un accès à la nourriture et aux soins, un éclairage intelligent peut répondre à la problématique de consommation d'énergie et de sécurité publique. Distribuer une eau conforme aux exigences sanitaires mais possédant également une qualité gustative irréprochable est l'un des services fondamentaux proposés par une ville à ses citoyens.

Néanmoins les villes s'agrandissent, afin d'être toujours en mesure de délivrer de l'eau

au nombre croissant d'usagers, et doivent faire face à plusieurs défis : gérer durablement des ressources hydriques limitées, assurer l'accès à l'eau potable aux populations non encore desservies et inciter les usagers à des comportements économes en eau. Ces défis sont d'autant plus cruciaux que les tensions sur les ressources en eau risquent d'être exacerbées sous les effets du changement climatique. L'augmentation de température et la baisse des précipitations prévues conduiraient, en effet, à la fois à une réduction des ressources et à une augmentation de la demande.

Un réseau d'alimentation en eau potable (AEP) intelligent doit alors permettre d'assurer la fourniture d'une eau saine et bonne à boire depuis son extraction aux différentes sources, jusqu'aux robinets des consommateurs, tout en maîtrisant l'impact sur la ressource tant hydraulique qu'énergétique.

1.2 Le petit cycle de l'eau

En France, ouvrir un robinet ou tirer la chasse d'eau sont des gestes de la vie quotidienne. Mais ces gestes anodins sont rendus possibles grâce à tout un processus impliquant de

(1) puiser l'eau dans le milieu naturel ;

(2) traiter l'eau pour la rendre potable ;

(3 & 4) transporter, stocker et distribuer l'eau traitée ;

(5) collecter et transporter les eaux usées ;

(6 & 7) dépolluer et rejeter l'eau dans la nature.

C'est le petit cycle de l'eau, aussi appelé cycle domestique de l'eau voir figure 1.1 ci-dessous.

FIGURE 1.1 – Le petit cycle de l'eau

Image source : Agence française pour la biodiversité - Réalisation Matthieu Nivesse, 2018

Contrairement à la circulation naturelle de l'eau sur Terre, le petit cycle de l'eau est artificiel. Il est assuré grâce à un ensemble d'infrastructures et d'acteurs ayant un rôle précis : pomper l'eau, la traiter, l'acheminer, etc.

La gestion du petit cycle de l'eau est assurée par les services publics d'eau et d'assainissement. De la responsabilité des collectivités locales, cette gestion couvre deux grandes missions : d'une part l'alimentation en eau potable, d'autre part l'assainissement des eaux usées.

La qualification de "cycle" peut être trompeuse : le petit cycle de l'eau est seulement une parenthèse dans le cycle naturel de l'eau. Ce n'est en aucun cas un cycle fermé, qui fonctionnerait dans une boucle continue : l'eau potable qui coule au robinet ne provient jamais des stations d'épuration.

Nous allons dans cette partie donner quelques informations concernant le fonctionnement des différentes étapes que constituent le petit cycle de l'eau. Depuis l'extraction et le traitement des eaux brutes, en passant par la distribution de l'eau traitées, jusqu'à la collecte des eaux usées et le rejet des eaux propres.

1.2.1 Puiser de l'eau dans le milieu naturel

L'eau potable est produite à partir d'eaux brutes prélevées dans le milieu naturel. En France, 60% de l'eau potable est produite à partir des eaux souterraines et 40% à partir des eaux de surface :

les eaux souterraines : elles sont prélevées par forage dans une nappe aquifère libre qui provient de l'infiltration des eaux de pluies dans un terrain perméable. Lorsque ces pluies rencontrent une couche imperméable, elles forment une nappe aquifère. La première nappe rencontrée est la nappe phréatique.

les eaux de surface : (cours d'eau, lacs, étangs) sont alimentées par des sources, le drainage des eaux souterraines et le ruissellement des eaux de pluie ; elles sont prélevées par captage au fil de l'eau, le plus souvent en amont de l'agglomération à desservir.

Le prélèvement des eaux de surface est géré de façon à concilier les débits du cours d'eau et les besoins des consommateurs. En effet, le débit du cours d'eau doit rester suffisant pour permettre la vie de la faune et de la flore aquatiques. L'exploitation des eaux souterraines se fonde sur l'étude de cartes hydrogéologiques qui localisent les gisements et représentent les niveaux d'eau. Des réseaux d'observation de la qualité des eaux, du débit des eaux de rivières et du niveau des nappes contribuent à cette gestion.

1.2.2 Traiter l'eau pour la rendre potable

L'eau prélevée dans les nappes d'eaux souterraines peut généralement être consommée sans traitement (à part une désinfection) si elle n'a pas été polluée par les activités humaines. En revanche, l'eau prélevée dans les rivières est impropre à la consommation. Chargée de débris de matières minérales (sable, limon...) ou organiques, elle est rarement limpide et saine. Il est nécessaire de la traiter pour la rendre potable. Ce traitement comprend généralement les étapes suivantes :

le dégrillage et le tamisage : l'eau passe à travers des grilles plus ou moins fines, afin d'éliminer les plus gros déchets (branches, feuilles mortes, sable et limon, débris divers etc.)

la clarification : elle permet de rendre l'eau limpide, en la débarrassant des matières en suspension, des algues et des particules colloïdales qu'elle contient. Elle s'effectue en deux temps :

1. on injecte d'abord dans l'eau un réactif chimique qui provoque la coagulation des particules. Les particules coagulées s'agglomèrent les unes aux autres et forment des flocons (c'est la floculation) ;
2. les flocons, plus lourds que l'eau, se déposent au fond du bassin de décantation et sont évacués régulièrement sous forme de boues.

la filtration : la clarification est améliorée en faisant passer l'eau à travers un lit de sable de 80 à 150 cm d'épaisseur : les particules encore présentes dans l'eau sont retenues au fur et à mesure de leur cheminement dans le filtre.

l'affinage : il sert à améliorer la qualité de l'eau et éliminer les goûts et les odeurs. Il comprend deux étapes :

1. l'ozonation qui facilite l'élimination des matières organiques et détruit les bactéries et les virus grâce au pouvoir désinfectant puissant de l'ozone,
2. la filtration sur charbon actif en grains qui élimine la matière organique qui pourrait générer des mauvais goûts et retient les polluants dissous dans l'eau comme les pesticides, hydrocarbures, etc..

Des ajustements physico-chimiques peuvent être nécessaires, notamment des corrections de l'acidité de l'eau pour éviter que celle-ci ne corrode les canalisations qui la transportent.

la désinfection c'est la dernière étape du processus. Elle permet d'éliminer les micro-organismes qui pourraient être dangereux pour la santé. Il existe plusieurs techniques de désinfection, mais la plus répandue est la chloration qui consiste à injecter dans l'eau du chlore selon un dosage précis. A la sortie de l'usine de traitement, le chlore est laissé en petite quantité dans l'eau, que l'on appelle chlore libre, pour protéger l'eau pendant son transport et le réseau de distribution contre le développement d'éventuels micro-organismes.

1.2.3 Transporter, stocker et distribuer l'eau traitée

Avant d'arriver au robinet des utilisateurs, l'eau potable emprunte un réseau souterrain de canalisation qui la conduit des usines de production d'eau potable jusqu'aux réservoirs. On distingue les conduites d'adductions qui sont destinées au transport des gros débits comme par exemple entre la station de traitement et les ouvrages de stockage (réservoirs) et le réseau de distribution qui assure la desserte vers tous les utilisateurs.

A la sortie de l'usine de production, l'eau potable est acheminée vers des réservoirs. Ceux-ci sont soit enterrés, soit surélevés (châteaux d'eau). La situation et la hauteur du château d'eau assure une pression suffisante dans tout le réseau et permet de distribuer l'eau par gravité à une pression régulière jusque dans les habitations. Les réservoirs constituent en outre une réserve d'eau potable pour les heures de consommation de pointe dans

une journée (le matin, à la mi-journée et le soir) et assurent une fonction de sécurité d'approvisionnement dans l'éventualité d'un incident sur les équipements d'alimentations du réseau ou d'un incendie. En France, en 2014, la consommation moyenne annuelle d'eau potable par habitant est de 52.2 m^3, soit 144.6 litres par jour.

La distribution de l'eau potable aux usagers se fait ensuite au moyen d'un réseau de canalisations qui relie les points de stockage aux lieux d'utilisation. Ce réseau peut être ramifié (structure en arbre c.f. Figure 1.2(b)) ou maillé (structure en treillis c.f. Figure 1.2 (a)). En 2013, en France, le réseau AEP est évalué à 996 000 kilomètres de conduites*, soit plus de 20 fois le tour de la Terre. Ces canalisations sont en général en fonte ou en plastique. Lorsque la population alimentée est importante, tous les ouvrages de production et de distribution sont surveillés en permanence grâce à un central informatique qui mesure le volume livré au réseau, les niveaux des réservoirs, la pression sur le réseau et la qualité de l'eau.

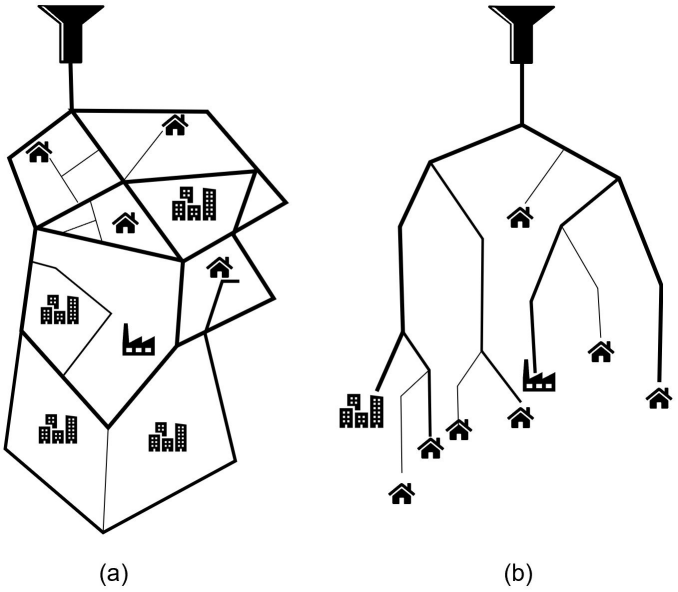


FIGURE 1.2 – Différentes architectures de réseaux : (a) maillé et (b) ramifié.

Toutefois, cette étape de transport de l'eau jusqu'au robinet de l'usager est source de pertes d'eau. Le vieillissement des réseaux conduit à l'apparition de fuites. En France, le rendement des réseaux (c'est-à-dire la quantité d'eau potable produite et effectivement distribuée) est estimé à 80%, c'est donc 20% de l'eau potable qui est perdue au cours de sa distribution (Rapport BIPE 2019). En moyenne pour 5 litres d'eau mis en distribution, 1 litre d'eau revient au milieu naturel sans passer par le consommateur. Les pertes par fuites représentent ainsi près d'un milliard de mètres cubes. Elles sont souvent dues à la vétusté des canalisations ou à une pression trop élevée, mais aussi aux mouvements des sols. On peut différencier les fuites selon trois états. 1) Les fuites diffuses, inconnues et non détectables par recherche de fuites, 2) les fuites détectables, inconnues mais détectables par recherche de fuites et 3) les casses manifestes, connues mais pas encore réparées [Lambert,

∗. Données : SISPEA (ONEMA) - 2013 / Source : Observatoire des services publics d'eau et d'assainissement - Panorama des services et de leur performance en 2013, 2016

1994].

1.2.4 Point de consommation d'eau

À la fin du 18e siècle les hygiénistes estimaient qu'une personne utilisait, pour l'ensemble de ses besoins, 15 à 20 litres d'eau. En France, au début du siècle dernier, la qualité de l'eau était médiocre et peu de villes disposaient de réseaux d'eau potable. Aller chercher l'eau à la source ou au puits, laver son linge au lavoir, toutes ces corvées faisaient partie des tâches quotidiennes, jusqu'à la 2nde Guerre Mondiale. Depuis chaque foyer est raccordé au réseau AEP par le biais d'un branchement. Ces tuyaux relient les canalisations jusqu'au compteur d'eau permettant de mesurer l'eau consommée. Cette mesure peut être faite manuellement par un technicien qui se déplace pour relever le volume consommé sur une échelle infra-annuelle ou automatiquement à distance via un émetteur positionné sur le compteur envoyant le volume consommé sur une échelle infra journalière.

Selon l'Onema la consommation moyenne journalière d'eau potable en France est de 145 litres par jour et par habitant. Pour obtenir l'utilisation globale d'une famille, il ne suffit pas de multiplier ce chiffre par le nombre d'individus vivant sous le même toit. Certains usages de l'eau impliquent en effet des utilisations identiques, quel que soit le nombre de personnes au foyer. Par exemple le nettoyage de l'habitat est un poste globalement incompressible, tout comme le volume d'eau nécessaire à une vaisselle ou une lessive en machine reste le même. Ces appareils tournent de la même façon pour une ou plusieurs personnes.

Ce chiffre diffère toutefois sensiblement en fonction d'un certain nombre de critères. Par exemple l'âge ; un adulte utilise nettement plus d'eau que les enfants (69 litres par jour) ou les personnes âgées (105 litres par jour). Le mode de vie aussi influe ; les sportifs se caractérisent par une consommation d'eau plus importante que la moyenne (204 litres par jour). Par ailleurs, le Français en vacances se montre moins économe, l'utilisation moyenne passe alors à 230 litres d'eau par jour. Le climat (région humide ou exposée à la sécheresse), la présence importante d'habitats individuels, l'existence de jardins, pelouses, piscines et l'activité touristique influent sur les besoins en eau et sur la consommation domestique moyenne d'eau du robinet par an et par habitant (109 litres par jour en région Nord-Pas-de-Calais contre 228 litres par jour en Provence-Alpes-Côte d'Azur).

Il est donc difficile de prévoir et estimer les volumes consommés en temps réel. Mieux comprendre les comportements et ainsi prévoir les consommations d'eau est un enjeu majeur. A la fois pour réduire l'impact du prélèvement sur certains milieux à risque mais aussi pour mieux dimensionner les services d'assainissement. En moyenne seulement 7% de l'eau que nous utilisons est dédiée à notre alimentation et 93% à l'hygiène et au nettoyage. La majorité de cette eau se retrouve dans le réseau d'assainissement.

1.2.5 Traiter et dépolluer les eaux usées

Une partie de l'eau une fois consommée se retrouve dans ce qu'on appelle le réseau d'assainissement. L'ensemble des eaux usées y sont collectées, ainsi que les eaux de pluie à travers un réseau unitaire ou séparatif séparant les eaux de pluie des eaux usées domestiques. L'eau est acheminée à travers des canalisations jusqu'à des stations d'épuration afin d'y être traitée et dépolluée. Elles mettent en œuvre des procédés artificiels qui imitent le

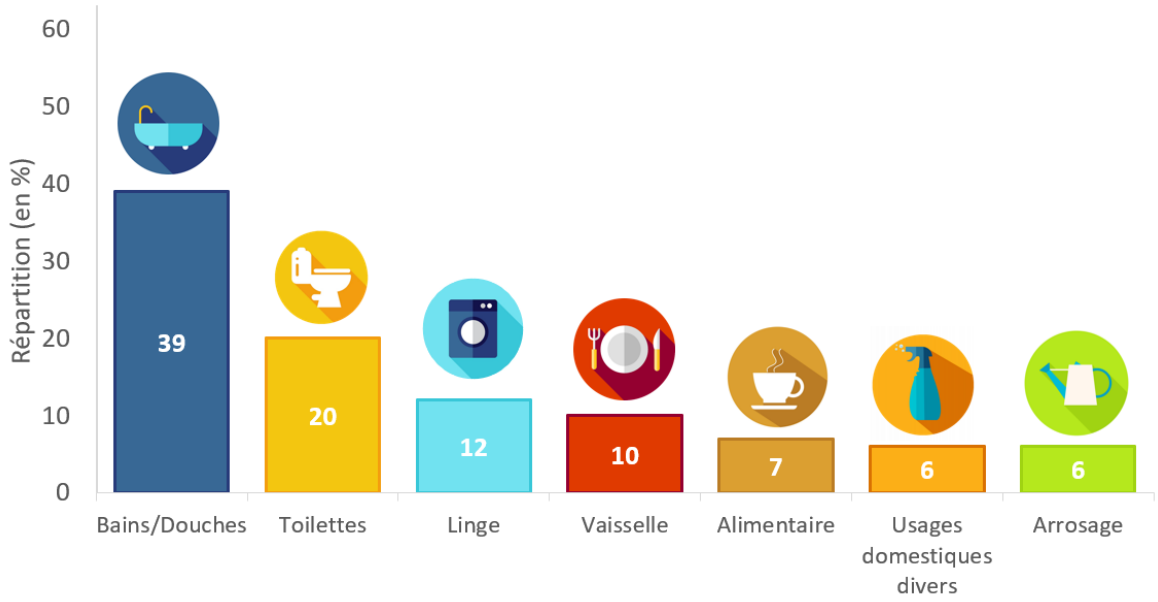


FIGURE 1.3 – Répartition de la consommation d'eau pour un ménage français (OMS 2018)

processus naturel d'auto-épuration des rivières. A la fin du traitement, la faible quantité de polluants rejetés permet de préserver la bonne qualité du milieu naturel.

Le réseau d'assainissement ne sera pas traité dans ce manuscrit. Cependant les travaux réalisés durant cette thèse peuvent être adaptés à celui-ci. Un cas d'application de parcours de graphe sur des données issues du réseau d'assainissement de la Métropole Bordelaise est présenté Section 2.4.3.a.

1.3 Réseau d'eau intelligent

Un réseau d'eau intelligent est un ensemble de solutions et de systèmes permettant aux opérateurs de réseaux d'eau de contrôler et diagnostiquer les problèmes, de prioriser et gérer, en continu et à distance, les opérations de maintenance et d'utiliser les données fournies pour optimiser tous les aspects de la performance des réseaux de distribution d'eau.

L'objectif est d'optimiser la performance du service de l'eau tout en améliorant l'efficacité énergétique, en préservant la ressource en eau et en surveillant la qualité de l'eau distribuée aux consommateurs.

À l'image des réseaux électriques et des réseaux de gaz intelligents, les réseaux d'eau deviennent intelligents grâce au déploiement des technologies de l'information et de la communication. Le déploiement de capteurs sur les réseaux permet de mieux comprendre (réaliser des mesures), d'optimiser (réaliser des analyses) et d'exploiter dans de meilleures conditions le réseau (notamment interagir avec le réseau).

Des capteurs, positionnés sur les usines de production dans le réseau de distribution ou au plus près du consommateur, permettent de surveiller et contrôler des paramètres clés liés à la qualité de l'eau (chlore, pH ou encore température) ou à la demande en eau

(consommations d'eau télé-relevées) et ceci sur un pas de temps fin ($\approx$ 15min). Mais au même titre que les villes grandissent, les réseaux de distribution grandissent eux aussi afin d'être toujours en mesure de délivrer de l'eau à ce nombre croissant d'usagers. Les réseaux étant de plus en plus maillés, afin de garantir une meilleure sécurisation de l'approvisionnement, cela complexifie énormément leurs fonctionnements hydrauliques et augmente les risques de stagnation dans certains tronçons.

Les informations fournies par les capteurs positionnés sur le réseau laissent donc forcément des zones d'ombre dans lesquelles il est de plus en plus difficile de maîtriser l'ensemble des phénomènes ayant lieu sur une maille géographique et temporelle fine.

L'interaction du chlore avec certaines matières organiques et inorganiques dans la phase liquide, les réactions avec le biofilm présent dans les conduites, la nature du matériau des conduites, le temps de séjour de l'eau, les interventions sur le réseau, les mélanges d'eau de qualités différentes, etc. sont autant de paramètres qui influent sur la cinétique de décroissance du chlore et donc potentiellement sur la qualité de l'eau.

De nombreux travaux de recherche fondés sur une approche déterministe de la décroissance du chlore en réseau et s'appuyant sur la connaissance des conditions hydrauliques ont déjà été réalisés [Vasconcelos et al., 1997]. Une thèse publiée en 2013 [Guépié, 2013] s'est intéressée à la construction de modèles de variation du chlore libre pour un réseau ramifié d'eau potable permettant de détecter une contamination dans le réseau suite à une variation anormale de la quantité de chlore. Ces modèles déterministes dépendant notamment du temps de séjour de l'eau dans les diverses sections du réseau et des cinétiques de décroissance des désinfectants sont établi en fonction de modèles hydrauliques du réseau de distribution.

D'autres travaux de recherche se basent eux sur des méthodes probabilistes ou d'apprentissage avec notamment des modèles de séries chronologiques [Karadirek et al., 2015], ou encore de régression généralisée dans des réseaux de neurones (General Regression Neural Network) [Bowden et al., 2006]. Ces modèles permettent la prédiction de la concentration de chlore en un point observé du réseau considérant les mesures en sortie d'usine et les temps de séjour sur quelques pas de temps. En revanche, ils ne permettent pas l'estimation des concentrations de chlore en tout point du réseau de distribution. De plus, l'application de ces modèles dans un système complexe comme un réseau maillé (soumis à des conditions hydrauliques difficilement appréhendables sur un pas de temps fin) génère des incertitudes liées au fait que toutes les composantes structurelles du réseau de distribution constituent des configurations non triviales et interagissent de manière assez complexes, rendant le calcul des temps de séjour ou des coefficients de cinétique de décroissance du chlore libre plus compliqué. De plus, les modèles hydrauliques (hors ligne) sont dans la majeure partie du temps utilisés à des fins de dimensionnement ou de planification [Danladi Bello et al., 2015] ou afin de diagnostiquer les conditions de fonctionnement des réseaux [Yoyo et al., 2016]. Dans ces modèles, des milliers de paramètres inconnus sont approximés à l'aide d'un échantillon à court terme d'un sous-ensemble de données hydrauliques issues de mesures ou de conditions particulières (jours de pointes, vacances etc.). La stabilité de ces modèles limite la fiabilité et l'efficacité d'un réseau de distribution d'eau potable, en particulier dans des conditions de défaillance partielle, car elle ne permet pas le réglage automatique des paramètres (courbes de débits, consommation d'eau, etc.) qui amélioreraient les performances. Ainsi, les résultats d'étalonnage peuvent ne pas représenter avec précision les conditions du système pour l'ensemble des

mesures. Des travaux récents encore en cours [Abu-Mahfouz et al., 2019, Abu-Mahfouz et al., 2016] ont proposé un cadre utilisant divers systèmes, méthodes et techniques avancés qui permettrait de développer un modèle hydraulique en temps réel pour la réduction des pertes en eau potable. Ces modèles hydrauliques temps réel nécessitent que l'ensemble des données y soient directement intégrées.

Ces incertitudes nous invitent à considérer l'environnement du réseau dans son ensemble, ainsi que les interactions entre les données mesurées, la typologie du réseau de distribution et les données ayant un impact sur le fonctionnement du réseau AEP.

Nous présentons dans la section suivante les données qui ont été exploitées dans le cadre de la thèse.

1.4 Données d'étude

Les données exploitées dans le cadre de cette thèse proviennent du contrat d'eau potable de Bordeaux Métropole (BM) dont la gestion a été déléguée à SUEZ. Depuis janvier 1992, Bordeaux Métropole a concédé le service de la distribution d'eau à SUEZ pour 23 des 28 communes du territoire. Les 5 communes non comprises dans le service de l'Eau de Bordeaux Métropole sont gérées par le syndicat d'alimentation de Carbon-Blanc. Dans l'ensemble du manuscrit on considérera deux réseaux AEP. Bien qu'ils soient tous les deux compris dans le contrat eau de BM, ils ne sont pas connectés entre eux. On notera *Bx* celui composé de 20 des 23 communes du contrat BM (en bleu Figure 1.4) et *Ambès* celui composé de 3 des 23 communes du contrat BM (en orange Figure 1.4).

Les travaux de recherche se sont concentrés sur la partie d'alimentation en eau potable (AEP) : une fois l'eau traitée, depuis les usines AEP jusqu'aux robinets des consommateurs. Nous présentons donc dans cette section les données générées autour de la gestion du réseau AEP du contrat de BM. Plus précisément celles nous permettant de constituer un graphe complexe représentant un réseau AEP complet. Il existe une grande diversité de données en fonction des paramètres clés du réseau qui sont étudiés : structure du réseau, qualité de l'eau, hydraulique du système, interventions, fuites, consommations d'eau etc. (voir Figure 1.5)

1.4.1 Squelette du réseau AEP - Données patrimoniales

Ces données concernent l'ensemble du patrimoine réseau de Bordeaux Métropole. Elles peuvent être directement observées depuis le Système d'Information Géographique (SIG) APIC. Elles sont constituées de plusieurs tables représentant chacune les différents objets physiques du patrimoine (réservoirs, canalisations, accessoires, branchements etc.). L'ensemble de ces objets sont géo-référencés et possèdent des caractéristiques qui sont retranscrites dans les tables (longueur de tuyaux, diamètre, date de pose, etc.). Ces données représentent le squelette du réseau. Au-delà de représenter le patrimoine enterré, elles ont un rôle central, celui de fournir une structure sur laquelle l'ensemble des données pourront être rattachées.

Les données patrimoniales sont stockées sous la forme d'une base de données SQL, consultable par le biais du SIG APIC. Il existe une table de données pour tous les accessoires et objets du réseau AEP représentés dans le SIG. Il est possible de faire des

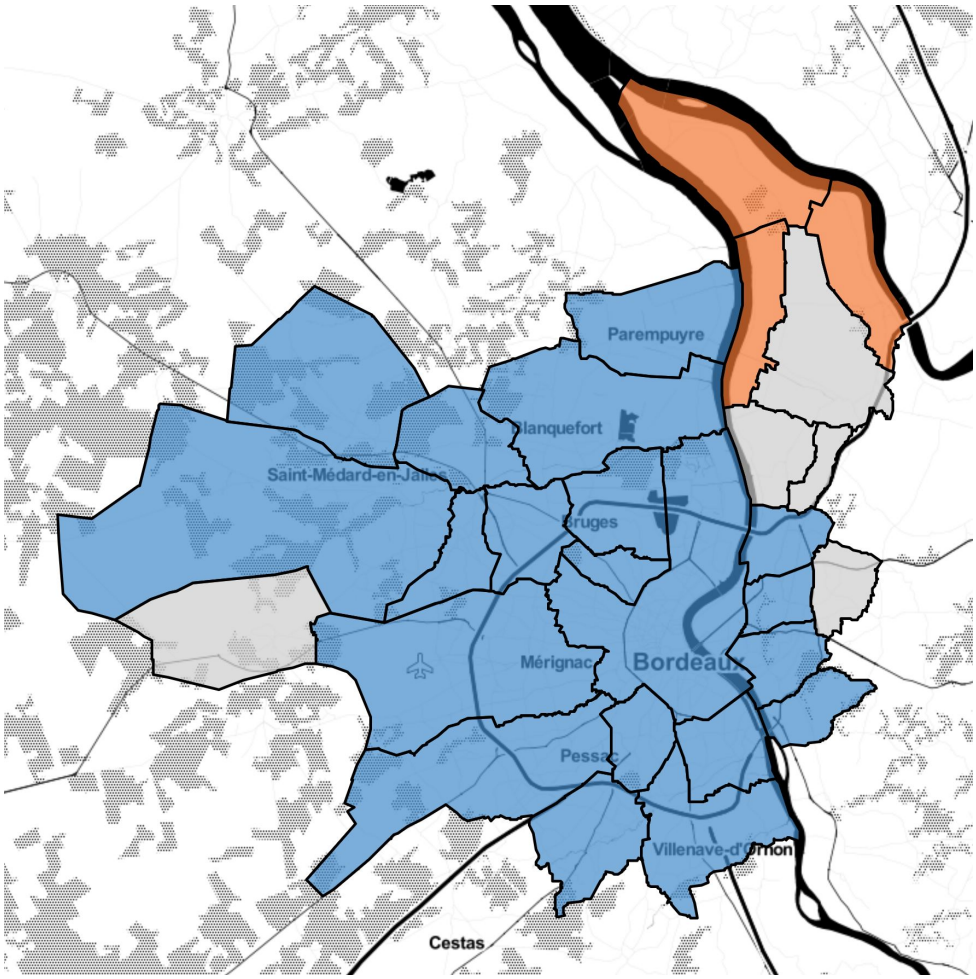


FIGURE 1.4 – Communes de la métropole bordelaise

extractions ponctuelles de ces tables. Le SIG du système d'eau potable est sans cesse mis à jour. Une grande partie des opérations effectuées sur le terrain (réparation, renouvellement, ajout, etc.) sont intégrées dans un outil de gestion des interventions, G2. Les modifications structurelles étant par la suite intégrées dans le SIG par un gestionnaire APIC. Les agents de terrain participent également à cette mise à jour lorsqu'ils rencontrent des écarts entre la réalité sur le terrain où ils interviennent et la configuration indiquée par le SIG. Ces mises à jour n'étant pas automatisées les délais entre une modification sur le terrain et son apparition dans le SIG peuvent varier. De ce fait, nous avons décidé d'effectuer une extraction de la base de données en fin d'année (décembre 2017) afin d'avoir un état structurel du réseau le plus récent et le plus à jour possible.

Le SIG n'a pas pour but de représenter de manière exhaustive l'ensemble des accessoires présents sur le réseau. Il existe différentes bases de données qui répertorient certains types d'objets. La base de données Pilier Patrimoine Visible (PPV) référence l'ensemble des équipements, comme les capteurs, présents en usine et sur le réseau de distribution.

Les caractéristiques structurelles du patrimoine

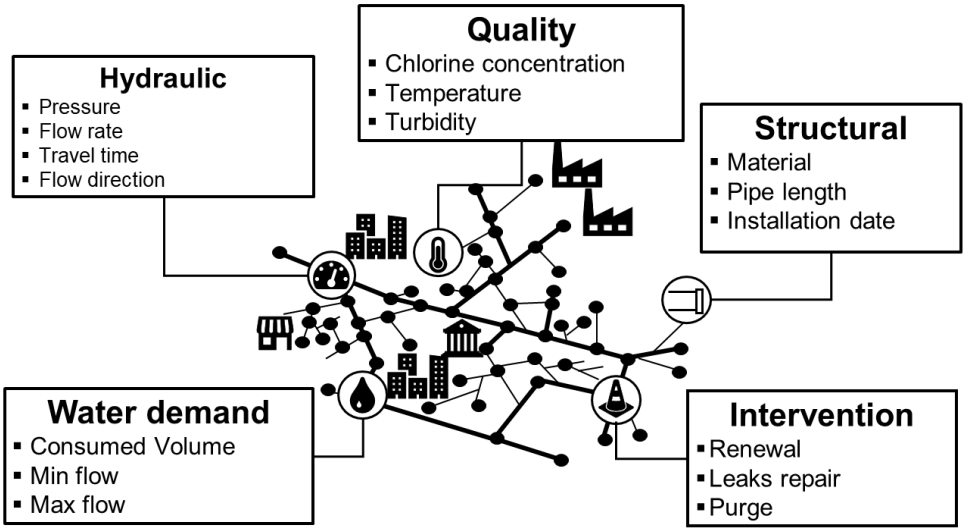


FIGURE 1.5 – Représentation schématique des données générées autour de la gestion d'un réseau AEP

a- Structure et topologie des réseaux AEP Dans le cadre du développement des réseaux AEP, différentes architectures de réseaux ont été réalisées. Elles différent par leur topologie, leur agencement et leur exploitation. Par topologie d'un réseau il faut comprendre l'ensemble de ses éléments, ses composants et leurs interconnexions dans le réseau. Dans la pratique la topologie du réseau est le résultat d'un compromis technico-économique. Le réseau AEP doit être en mesure de répondre aux objectifs d'alimentation en eau potable visés, tout en prenant en comptes les contraintes physiques et techniques liées au territoire sur lequel il est installé.

Il est courant de rencontrer deux architectures différentes de réseau (Figure 1.2) :

— les réseaux dit *maillés*. Ils sont principalement installés pour raccorder les usagers en zone urbaine. Ils garantissent une meilleure sécurisation de l'approvisionnement car l'eau potable peut arriver chez l'usager en suivant plusieurs chemins. Néanmoins des points de stagnation peuvent apparaître dans certains tronçons. La stagnation augmente le risque de prolifération bactérienne en cas de défaillance de la chloration.

— les réseaux dit *ramifiés*. Ils sont principalement installés en zone rurale où il est nécessaire de couvrir de longues distances pour desservir les usagers. L'eau circule toujours de façon unilatérale dans les ramifications réduisant ainsi le risque de prolifération bactérienne. Néanmoins en cas de coupure d'un tronçon, toutes les ramifications "filles" sont privées d'eau. La sécurisation de la desserte en eau est donc moins grande que pour un réseau maillé.

Le réseau considéré ici est un vaste réseau de distribution, couvrant plusieurs communes. La topologie de celui-ci est donc hybride. Certains secteurs étant très urbanisés nécessitent une topologie maillée et d'autre beaucoup plus ruraux impliquent une structure ramifiée.

Le territoire de Bordeaux Métropole présente une grande amplitude de relief. Ainsi, d'Ouest en Est, les terrains situés sur le secteur de Saint-Médard-en-Jalles (+40/+60 m

NGF[†]) descendent avec une faible pente vers la Garonne (+2/+6 m NGF). A l'inverse de la rive gauche de la Garonne, les terrains situés sur la rive droite montent en pente franche jusqu'à +50/+80 m NGF.

Cette variation de niveau sur le territoire implique des variations de pressions sur le réseau qui doivent être maîtrisées. Les points les plus bas ayant les pressions les plus élevées. Ainsi le réseau est divisé en 15 étages de distribution principaux et chacun possède une référence de pression, qui lui est propre. En situation normale, chaque étage de pression est isolé des autres et possède ses propres ressources en eau et stockage. En situation anormale, des interconnexions permettent d'alimenter un étage par un autre. Il est possible d'imposer la pression dans un réseau d'eau de 2 façons. La modulation de pression à l'aide de vannes permettant de faire varier la pression dans le réseau par laps de temps d'une heure. Et la régulation de pression qui se fait par la variation du débit injecté dans le réseau à l'aide de pompes à vitesse variable.

b - Les canalisations Les canalisations sont les principales conduites de distribution d'eau potable, elles acheminent l'eau du point de stockage jusqu'aux branchements. Il existe plus d'une centaine de caractéristiques recensées pour les canalisations dans la base de données (MMS)-APIC. Allant de paramètres graphiques pour l'affichage dans le SIG, jusqu'à la retranscription de caractéristiques physiques de chacune des canalisations, en passant par des informations contextuelles pouvant être rattachées à des canalisations. Plus de 67 000 canalisations sont référencées dans cette table. Cette liste représente l'ensemble des canalisations du réseau de la métropole bordelaise.

Comme indiqué précédemment, nous nous concentrons sur la partie distribution en eau potable une fois l'eau traitée, en aval des stations de traitement. Ceci signifie que nous ne prenons pas en compte les canalisations positionnées dans le réseau d'adduction en amont des usines de production d'eau potable.

Une fois filtré pour ne prendre en compte que les canalisations du réseau AEP nous dénombrons 64 783 canalisations. Sur l'ensemble de ces canalisations, 63 595 font partie du réseau *Bx* pour un linéaire de plus de 2 900 Km et 1 188 du réseau *Ambès* pour un linéaire d'un peu plus de 100 Km. Il s'agit d'ouvrages dont la longueur graphique dans le SIG peut varier entre quelques centimètres pour les plus petites et quelques centaines de mètres pour les plus grandes Annexe 1.6.

Les canalisations d'eau potable ne sont pas toutes faites du même matériau, celui-ci ayant évolué avec le temps. Au milieu du XIX$^{\text{ème}}$ siècle, les réseaux étaient principalement faits d'acier ou de fonte appelée fonte grise alors qu'au début du XX$^{\text{ème}}$, on voit apparaître des conduites en béton ou en amiante ciment. Une nouvelle génération de fonte, la fonte ductile, est ensuite apparue dans les années 60. Enfin dans les années 50, les premières conduites en polychlorure de vinyle (PVC) sont installées sur le réseau de distribution. La Figure 1.7 illustre la répartition des matériaux en fonction du linéaire de réseau.

Le réseau AEP de la métropole bordelaise est majoritairement constitué de canalisation en fonte, représentant plus de 80% du linéaire du réseau *Bx* (50% de fonte grise et 30% de fonte ductile) et près de 60% de celui d'*Ambès* (35% de fonte grise et 25% de fonte

†. Nivellement général de la France (NGF), constitue un réseau de repères altimétriques disséminés sur le territoire français métropolitain continental. +1m NGF représente un mètre au dessus du repère NGF

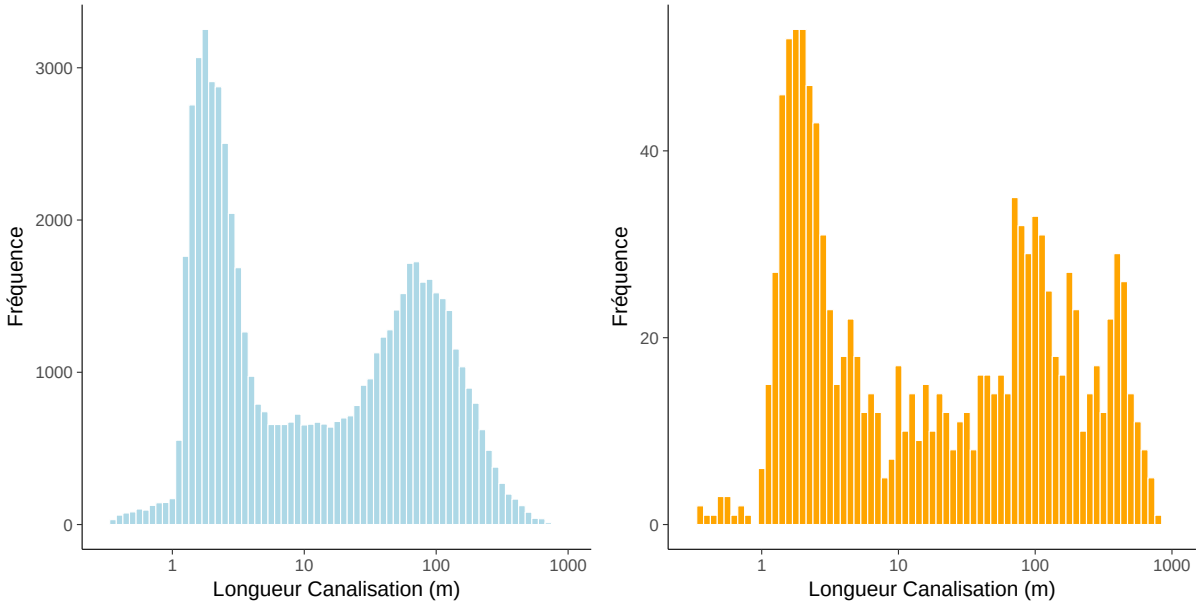


FIGURE 1.6 – Longueur des canalisations pour les réseaux *Bx* (à gauche) et *Ambès* (à droite)

ductile). Le PVC est de plus en plus utilisé lors des renouvellements pour remplacer les canalisations en acier, amiante ciment et béton. Une faible proportion des canalisations ont un matériau inconnu qui n'est pas encore répertorié dans la base de donnée SIG. Seulement 900m de canalisation est encore inconnu sur le réseau *Ambès* et approximativement 4 km pour celui de *Bx*. Ces canalisations sont pour la plupart des canalisations posées dans la fin du XX$^{\text{ème}}$.

Le diamètre des canalisations varie de quelques centimètres, 2.5cm pour les plus petites, jusqu'à 1 m de diamètre pour les canalisations de transport, induite par la pression et la quantité d'eau nécessaire pour répondre à la demande. (Figure 1.8).

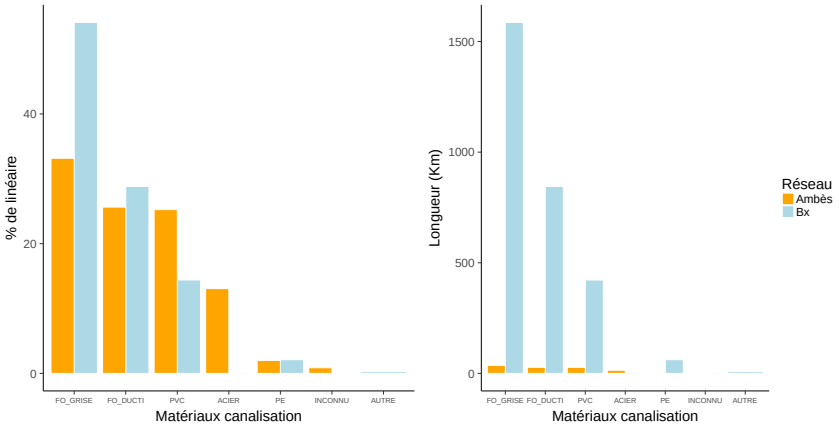


FIGURE 1.7 – Linéaire de réseau par classe de matériaux de canalisation pour les réseaux *Bx* en bleu et *Ambès* en orange

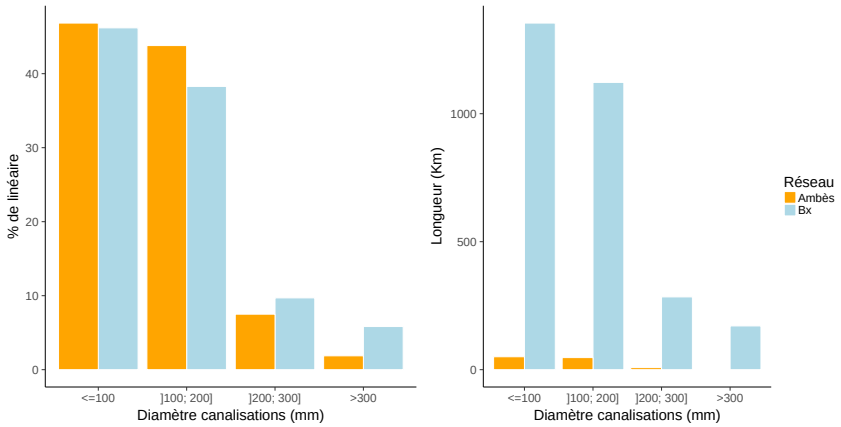


FIGURE 1.8 – Linéaire de réseau par classe de diamètres pour les réseaux *Bx* en bleu et *Ambès* en orange

c - Les branchements Les branchements, raccordés aux canalisations, sont les tuyaux par lesquels l'eau sera conduite depuis le réseau public vers le domaine privé des usagers. Un compteur séparant le patrimoine canalisation du patrimoine branchement.

Ils sont plus nombreux que les canalisations, mais d'un linéaire plus faible et de plus petits diamètres. Le réseau de distribution comporte environ 190 000 branchements particuliers alimentant plus de 270 000 points de livraison avec compteurs. L'écart entre le nombre de branchements et le nombre de points de livraison avec compteurs s'explique par l'existence de compteurs divisionnaires (en habitat collectif), configuration où un branchement physique unique assure l'alimentation de plusieurs points de livraison. Plus de 8 000 poteaux incendie (hydrants) complètent ce premier dénombrement de branchements physiques.

Chaque branchement est raccordé à un compteur permettant à la fois de quantifier les volumes d'eau consommé mais aussi pour référencer l'ensemble des usagers et clients du réseau de distribution d'eau potable. On peut donc attribuer à chaque branchement un point de livraison à partir de la base de données clientèle.

La base clientèle étant celle de référence pour le SIG, l'écart entre les 188 200 branchements cités ci-dessus et les 216 890 branchements avec une référence clientèle recensés dans le SIG (APIC) s'explique par les branchements en attente de référencement et ceux en attente de suppression.

d - Les accessoires On regroupe dans cette catégorie de données l'ensemble des "objets" pouvant relier deux canalisations entre elles. Ce lien est dans la plupart du temps décrit dans le SIG par des jointures de tables entre les canalisations et les tables de ces dits accessoires. On considérera trois types d'accessoires : les raccords, les vannes et les appareils de mesures.

Les raccords peuvent être de trois types selon le nombre de canalisations raccordée. Simple pour raccorder deux canalisations, en T pour trois et en croix pour 4. Il existe certains raccords présents dans le SIG qui ne sont pas présents sur le terrain, ils sont ajoutés lors du dessin de certaines canalisations complexes. On considérera dans l'ensemble du manuscrit la totalité des raccords étant présents dans le SIG.

e - Les usines de productions et réservoirs Les usines de productions ainsi que les réservoirs font eux aussi partie des accessoires du réseau AEP. Ils connectent des canalisations entre elles. Les usines de production d'eau potable séparent le réseau d'adduction du réseau de distribution.

f - Modulation de pression/secteur hydraulique La sectorisation est un moyen pour maîtriser les pertes en distribution et optimiser la recherche active des fuites. Pour des réseaux de distribution AEP de grande taille, comme celui de la métropole bordelaise, elle constitue un outil de diagnostic parfaitement adapté. Elle consiste à identifier plusieurs sous-réseaux ou secteurs, dans lesquels les volumes mis en distribution sont mesurés en permanence ou de façon temporaire. De cette manière, il est possible de calculer des indicateurs caractéristiques de la performance hydraulique.

La totalité du réseau de distribution de BM est divisé en secteurs de niveau I (linéaire de réseau supérieur à 100 km) et de niveau II (linéaire de réseau inférieur ou égal à 100 km). A fin 2017, on comptabilise :

— 10 grands secteurs de Niveau I, homogènes en termes de qualité d'eau distribuée dans ces zones (délimités en jaune Annexe D),

— 55 secteurs de Niveau II, avec un linéaire de réseau de 3 km pour le plus petit à 110 km pour le plus grand (délimités en orange Annexe D).

1.4.2 Données métrologiques

Dans cette section nous allons nous concentrer sur la présentation des données métrologiques. C'est à dire l'ensemble des données concernant le suivi de paramètres clés liés à la gestion des réseaux de distributions AEP, issues de capteurs positionnés sur le réseau. Ces paramètres sont mesurés sur des pas de temps fins allant de quelques minutes pour le suivi des flux ou de la qualité de l'eau, jusqu'à des mesures sur un pas de temps horaire pour les volumes consommés.

1.4.2.a Descriptions de l'environnement de données

Il existe plusieurs bases de données en fonction des paramètres mesurés sur le réseau de distribution. La plupart d'entre-elles sont tout d'abord surveillées en temps réel par le centre de télé-contrôle AUSONE[‡]. Organe de gestion technique centralisé, il permet de piloter et surveiller 24h/24h l'ensemble des installations assurant l'alimentation en eau potable du réseau de la métropole bordelaise.

L'ensemble de ces données sont ensuite stockées dans le logiciel Aquacalc. Il permet l'archivage de données acquises ou calculées. Les valeurs sont stockées dans des fichiers propres à Aquacalc ou dans toute base de données compatible ODBC (Access, Oracle, etc.). La pérennité de l'archivage peut être de plusieurs années. La consultation des données peut se faire sous forme de courbes ou de tableaux. Le module d'accès web permet à tout utilisateur du réseau d'entreprise d'accéder facilement aux données et aux rapports à l'aide d'un navigateur. Les données peuvent être ensuite extraites, sous forme de bilan en spécifiant les données et les périodes voulues.

‡. Automatisation des Unités de Surveillance et d'Optimisation des Nappes et de l'Eau

Le suivi des consommations d'eau est divisé en deux bases de données. Odyssée qui intègre toutes les fonctionnalités métier pour la gestion de la relation client, de la facturation et des points de livraison. La totalité des compteurs y sont répertoriés. On peut retrouver des informations relatives à chaque client (adresse, coordonnées, etc.) et chaque compteur (volumes consommés, diamètre du branchement, date de pose, etc.). La base de données SITR quant à elle regroupe les mesures des compteurs équipés d'émetteur télé-relevé (TLRV).

1.4.2.b Suivi des flux

Le suivi des flux concerne les capteurs permettant de surveiller les installations assurant l'alimentation en eau potable. Il va s'agir de débitmètres mesurant les quantités d'eau transitant en des points précis (en m^3/h), les capteurs de pression, des capteurs de niveau mesurant les quantités d'eau présentes dans les cuves ou des compteurs calculant les volumes d'eau consommés au point de livraison. Au delà de fournir des indications sur le bon fonctionnement hydraulique du réseau AEP, ces données nous fournissent des informations sur les signaux présents dans le réseau.

Nous allons nous concentrer ici sur les capteurs pouvant nous fournir des informations sur les flux transitant dans le réseau AEP, depuis les usines de production jusqu'aux points de livraisons. Nous avons pour cela catégorisé 3 types de flux,

1. les flux entrants dans le réseau AEP ou appelés volumes livrés au réseau (VLAR),
2. les flux transitant, représentant la mesure d'un signal de débit en des points particuliers du réseau et
3. les flux sortants qui sont mesurés à l'aide des compteurs positionnés sur les branchements.

Les volumes livrés au réseau (VLAR)

Les VLAR représentent les volumes mis en distribution. Ils sont mesurés par des débitmètres positionnés sur chacune des usines de production d'eau potable et sur les réservoirs.

On dénombre 38 usines de production d'eau potable et 22 réservoirs sur le réseau AEP de BM. C'est donc 60 points sur lesquels on observe un signal de débit sur un pas de temps 6 minutes. Toutes les usines de production ne fonctionnent pas en même temps ni même en continue, comme on peut le voir avec la Figure 1.9. La courbe verte représente le débit sortant de l'usine de production de Tremblay (situé au nord de la commune de Tremblay), et la courbe violette celle en sortie du réservoir de Beauregard. Ces deux courbes sont très cycliques avec un cycle de 24h, représentant leurs périodes de fonctionnement journalier. On remarque que la courbe d'usine possède un plateau tous les jours de 22h00 à 06h00 qui représente la période pendant laquelle l'usine ne livre plus d'eau potable. En revanche le signal du réservoir possède lui aussi un cycle journalier mais beaucoup moins marqué. En effet les réservoirs distribuent l'eau par gravité à une pression régulière mais fortement dépendante des consommations et des modes de fonctionnement des usines de production. Par exemple lorsque les VLAR des usines sont nuls les réservoirs peuvent prendre le relais et continuer d'alimenter le réseau en eau potable. Cela se traduit sur la Figure 1.9 par le pique de débit que l'on observer sur la courbe violette tous les jours à partir de minuit.

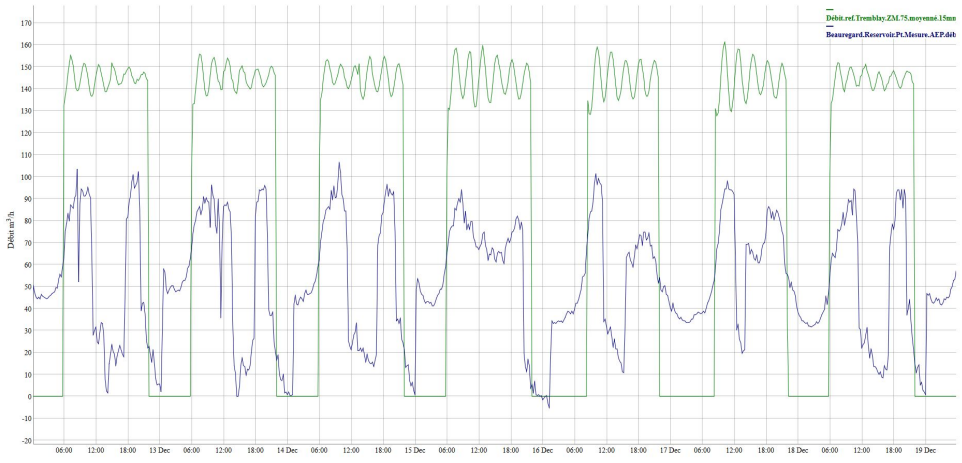


FIGURE 1.9 – Courbe de débits en m^3/h sur une semaine pour un réservoir (en violet) et une usine de production (en vert)

Le volume mis en distribution et les usines employées pour répondre à la demande vont dépendre de plusieurs critères. Ils vont dépendre de la période de l'année, de la météo et de la disponibilité de la ressource. Par exemple, une année qui s'est avérée particulièrement sèche avec une baisse significative de la pluviométrie limitant la recharge hivernale, et donc la capacité de mise-en-réseau, et des températures élevées tout au long de l'année va fortement impacter le mode de fonctionnent des usines de production. Toute cette optimisation sur les temps de fonctionnement des usines de production et de remplissage des réservoirs est effectuée par les exploitants et est suivie et pilotée en temps réel avec le télé-contrôleur AUSONE.

Volumes transitant

Le suivi des VLAR est un indicateur permettant d'évaluer l'évolution de la demande en eau sur un territoire. Néanmoins mesuré à l'échelle globale d'un réseau de distribution il ne permet pas à lui seul de surveiller et contrôler efficacement les volumes transitant et notamment les pertes potentielles en eau liées aux fuites. C'est pourquoi, comme expliqué Section 1.4.1, la sectorisation du réseau AEP de BM, en deux niveaux, a été mise en place. En bloquant le flux en certains points à l'aide de vannes de sectorisation et en positionnant des débitmètres en certains points du réseau on est capable de mesurer les volumes entrants et sortants de ces secteurs.

Il existe plus de 100 débitmètres de sectorisation positionnés sur le réseau mesurant un débit sur un pas de temps de 15 minutes. Le débit mesuré peut être décomposé en deux valeurs en fonction du sens de lecture du débitmètre. Le débit de *voie 1* qui est le débit affiché en valeur positive par le débitmètre et le débit de *voie 2* en valeur négative comme l'illustre la Figure 1.10. Ainsi connaissant le positionnement physique du compteur sur le réseau, on est capable de déterminer localement le sens d'écoulement du flux mesuré.

Volumes sortants - consommations Les volumes sortants regroupent l'ensemble des flux quittant un secteur ou le réseau par différents biais. Que ce soit le remplissage d'un réservoir, la vente d'eau à d'autres communes ou encore les volumes consommés par les

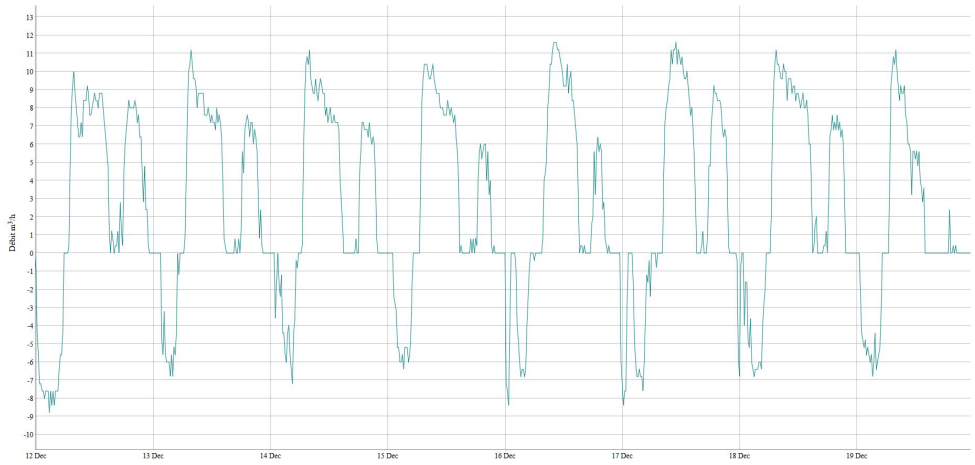


FIGURE 1.10 – Courbe de débits en m^3/h sur une semaine pour un débitmètre de sectorisation, les valeurs positives indiquant un débit en voie 1 et les valeurs négatives indiquant un débit en voie 2

usagers. C'est le dernier chaînon du suivi des flux.

Afin de suivre et appréhender la demande en eau chaque point de service (lieu où l'eau quitte le domaine public) est équipé d'un compteur mesurant les volumes sortants. Ce volume est appelé index et comptabilise la totalité du volume ayant transité à travers le compteur. Ces compteurs sont initialement installés à des fins de facturation, une fois les index relevés. La facturation est faite deux fois par an. Les index sont quant à eux relevés par des opérateurs sur le terrain une fois par an. Cette relève est effectuée par secteur et permet au service clientèle de mettre à jour les données de volumes consommés et ainsi ajuster la facturation par rapport à celle ayant été faite sur une estimation.

Afin d'améliorer la facturation de la consommation et de mieux appréhender la consommation des usagers sur des pas de temps plus fins, certains compteurs ont été équipés de transmetteur télé-relevé (TLRV). Ces compteurs TLRV transmettent tous les jours entre 4 et 24 index permettant d'obtenir les volumes consommés sur un pas de temps infra-journalier ou horaire. Malheureusement cette technologie est très onéreuse il apparaît alors d'un point de vue économique compliqué d'équiper l'ensemble des 250 000 compteurs avec la TLRV. Les compteurs étant ainsi équipés de cette technologie ont été sélectionnés stratégiquement. Premièrement ceux que l'on appelle les gros consommateurs, ils ont des typologies de consommation inhabituelles et consomment des volumes très importants d'eau. De manière générale ces compteurs représentent des entreprises privées utilisant l'eau potable comme matière première dans leurs cycles de production.

Les informations fournies par les compteurs TLRV permettent de définir des profils de consommation en fonction de deux critères : le diamètre du branchement et le type d'usager.

Les estimations des consommations non TLRV ne prennent donc en compte que l'historique de consommations TLRV des groupes de consommation auxquels elles appartiennent et leur propres historique de relevé d'index. Si on souhaite obtenir une informations plus proche de la réalité sur des pas horaires il faudrait être en mesure de prendre en compte à la fois les volumes entrants mais aussi les volumes transitant à chaque instant pour les ajuster. Nous présentons Section 2.4.3.b une méthode permettant d'ajuster les estimations

de consommation en fonction des flux observés temps par temps à l'aide de la structure de données présentée ci-après.

1.4.2.c Suivi de la qualité

Les dernières données métrologiques concernent le suivi qualité de l'eau potable, depuis sa mise en distribution jusqu'aux points de livraison. Dans un contexte sécuritaire, le risque d'atteinte ou de contamination volontaire de l'eau potable est à prendre en compte et à anticiper. La mise en œuvre de protection sur les sites eau potable, associée à la maîtrise de procédures de sûreté des sites permettent de retarder et d'alerter en cas d'intrusion.

Sur le réseau de distribution ce sont les capteurs qualité associés au pouvoir désinfectant du chlore qui permettent de suivre et d'être alerté en cas de dérive de la qualité de l'eau distribuée. Ces capteurs mesurent les concentrations de chlore, le pH et la température sur un pas de temps de 15 minutes.

De nouveaux capteurs sont à l'étude, capables de suivre des paramètres cibles de l'évolution de la qualité de l'eau tels que chlore, pH, température, conductivité, turbidité, UV et COT. L'introduction de ces nouveaux paramètres permet d'élargir le spectre de polluants détectables et d'anticiper plus efficacement l'évolution de la qualité de l'eau en réseau. Néanmoins, lors de ces travaux nous nous sommes concentrés sur le suivi des concentrations de chlore étant donné le peu de capteurs nouvelle génération mis en service.

Les premières installations de capteurs qualité sur le réseau de BM ont été réalisées au début des années 2000. A cette époque, la logique de déploiement était basée sur l'unité qualité eau du contrôle réglementaire. C'est à dire s'assurer de la conformité de l'eau mise en distribution.

La connaissance de la qualité de l'eau et des zones d'influence des multiples stations alimentant le réseau de distribution, a conduit à proposer courant 2015, un programme pluriannuel visant à densifier le parc de capteurs qualité réseau. Ce programme est quant à lui basé sur la mise en place d'un capteur par zone de qualité d'eau différente.

Détecter, localiser et suivre la dérive de la qualité de l'eau est une première étape. Comme imposé par la norme ISO 22000 [§], il apparaît nécessaire et indispensable de pouvoir limiter la propagation et retirer de la distribution une eau potentiellement non conforme. Cette notion met en exergue toute la difficulté qui réside dans le fait que le système de production d'eau potable soit un processus continu.

35 capteurs positionnés sur le réseau, à cela s'ajoute les capteurs positionnés sur 48 usines de production et les 16 réservoirs.

1.5 Conclusion du chapitre

Un des premiers enjeux permettant de prendre en compte toute l'inter-connectivité et l'interdépendance associée au réseau de distribution d'eau potable est de représenter les données sous la forme d'un réseau de graphe. Depuis l'émergence des réseaux sociaux beaucoup de travaux et de recherche dans le domaine de l'analyse des graphes ou de la théorie des graphes ont été effectués. La grande majorité de ces travaux se basent au

§. La norme ISO 22000 est une norme internationale, relative à la sécurité des denrées alimentaires.

même titre que l'analyse des réseaux sociaux sur l'analyse de communauté qui permettent d'obtenir des informations sur la topologie du graphe afin par exemple de détecter les points vulnérables du réseau de distribution d'eau potable [Wang et al., 2012], ou encore de déterminer la position optimale de poste de re-chloration dans un réseau de distribution d'eau potable [Said et al., 2016].

En conclusion, les travaux issus de la littérature nous ont permis de constater que des modèles de prédiction étaient déjà existants mais ne répondaient pas au problème d'estimation des paramètres d'exploitation (tel que le chlore) en tout point du réseau et en temps réel. En revanche les avancées scientifiques sur l'analyse des réseaux de graphes nous invitent à transposer des méthodes d'estimations existantes au cas particulier d'un réseau de graphe. L'approche choisie est pluridisciplinaire en croisant à la fois des méthodes informatiques, statistiques et des méthodes empruntées au Big Data. La première phase consiste en la mise en place d'une architecture des données capable de prendre en compte l'ensemble des données hétérogènes pouvant être générées autour du contrôle et de la surveillance des réseaux AEP présentés dans la section précédente. La seconde phase consiste en l'application de méthodes afin d'être capable de répondre à la problématique d'estimation en tout point et tout instant de paramètres clés liés à la gestion d'un réseau AEP.

Chapitre 2

Théorie des graphes et réseaux d'alimentation en eau potable

Ce chapitre a pour but d'introduire le concept de graphes mathématiques et de données structurées sous la forme d'un graphe. Nous présentons dans un premier temps les notations ainsi que quelques notions de base issues de la théorie des graphes.

Nous décrivons ensuite les données structurelles sélectionnées permettant la création des nœuds et arêtes du graphe de terrain représentant un réseau AEP, ainsi que les données métrologiques pouvant être utilisées pour labéliser les éléments du réseaux. Nous analysons ensuite les propriétés structurelles des graphes afin de caractériser un graphe et les connexions entre ses nœuds. Une autre approche d'analyse de graphe associant la connectivité d'un graphe à l'analyse propre de certaines matrices telles que la matrice d'adjacence ou la matrice Laplacienne est aussi présentée.

Enfin nous présenterons quelques concepts et algorithmes existants autour des réseaux AEP structurés sous la forme d'un graphe, ainsi que des applications réalisées à partir de cette structure de données pour répondre à des problématiques métiers.

2.1 Théorie des graphes et réseaux d'alimentation en eau potable

Les bases de l'analyse des réseaux dans les sciences, en particulier le fondement mathématique de la théorie des graphes, sont souvent placées dans la solution d'Euler de 1735 au problème désormais célèbre du pont de Königsberg, dans laquelle il prouvait qu'il était impossible de parcourir les sept ponts de cette ville de manière à ne les traverser qu'une seule fois.

Les fondements mathématiques formels de la théorie des graphes sont posés dans les années 1800, D. König étant cité comme l'un des premiers architectes clés. C. Berge, R. Faure et A. Kaufmann en France, F. Harary, Ford Lester Randolph Jr. aux États-Unis, le polonais K. Kuratowski et le hongrois P. Erdös, sont également considérés comme les précurseurs du développement de la théorie des graphes et de ses applications au $20^{\text{ème}}$ siècle.

Les graphes peuvent être de diverses formes et tailles, mais il existe un certain nombre de types de graphe couramment rencontrés dans la pratique. Ces graphes sont regroupés

par famille ayant des topologies structurelles similaires, décrivant la manière dont leurs nœuds et arêtes sont organisés. Il est intéressant de connaître ces différentes familles car elles possèdent des propriétés spécifiques permettant des analyses particulières.

On pourra citer par exemple les graphes *homogènes* ou *réguliers* reproduisant des schémas particuliers, les graphes *hiérarchiques* pouvant être découpés en niveaux ou couches hiérarchiques, les graphes *polarisés* qui eux sont tels que tous les sommets sont rattachés à un seul et même sommet central. Ces graphes sont considérés comme des objets mathématiques et ont été largement étudiés dans la littérature du fait de leurs caractéristiques. Ils ont permis notamment l'élaboration d'algorithmes optimisés en fonction de la nature du graphe.

Dans ces travaux nous nous concentrerons sur un type de graphe appelé graphe complexe dans le sens où leur structure est induite par leur nature technologique. Ces graphes sont considérés comme non triviaux car leurs formes ne suivent pas forcément un schéma particulier permettant de les faire figurer dans des familles déjà étudiées. On peut prendre l'exemple des réseaux de communication (téléphone, internet, etc.), réseaux de transport (routes, lignes aériennes, etc.) et des réseaux d'énergie (réseaux électriques, de gaz). La topologie de ces réseaux évolue en fonction des objectifs et des optimisations nécessaires. Par exemple la structure des réseaux d'eau potable évolue en fonction des modifications de l'architecture de la ville et sa topologie est induite par des contraintes technico-économiques.

Au cours des dernières décennies, l'étude des réseaux complexes s'est développée rapidement et a attiré de nombreux chercheurs de différentes disciplines et domaines d'application. Les méthodes et outils de la théorie des réseaux complexes fournissent une capacité analytique permettant de concevoir, d'optimiser, d'exploiter et de maintenir ces types de réseaux. On peut citer par exemple, ([Carreras et al., 2004, Zio and Golea, 2012a, Alipour et al., 2013, Saniee Monfared et al., 2014]) pour les réseaux électriques, ([Buckwalter, 2001, Shen and Gao, 2008, Bagler, 2008, Caschili and De Montis, 2013]) pour les réseaux de transport et ([Yazdani and Jeffrey, 2011, Shuang et al., 2014, Liu et al., 2015, Di Nardo et al., 2017, Di Nardo et al., 2018]) pour les réseaux d'eau potable.

2.2 Notations et définitions

Cette section a pour objectif d'introduire les notations qui seront utilisées tout au long de ce manuscrit ainsi que quelques définitions et concepts classiques de la théorie des graphes, tels qu'ils peuvent être trouvés dans des ouvrages de référence ([Gondran and Minoux, 1995, Newman, 2003, Kolaczyk and Csardi, 2014, Barabási, 2016]).

2.2.1 Définitions élémentaires

Graphe

Un *graphe* G est défini comme un couple $G = (V, E)$ composé d'un ensemble de *sommets* $V = \{1, ..., N_v\}$ (*vertices* en anglais) et d'un ensemble d'*arêtes* E (*edges* en anglais) reliant certaines paires de sommets de V. Formellement, $E \subseteq V \times V$. Nous considérons le cas de graphes finis, c'est à dire ayant un nombre fini de sommets et d'arêtes et notons $N_v = |V|$ et $N_e = |E|$.

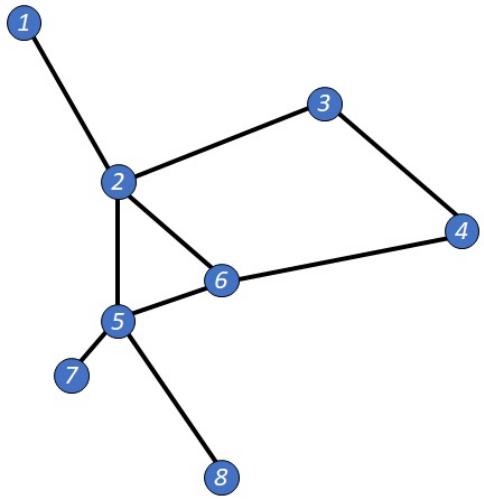


FIGURE 2.1 – Un graphe planaire $G = (V, E)$ avec l'ensemble de sommets $V = \{1, 2, 3, 4, 5, 6, 7, 8\}$ et l'ensemble d'arêtes $E = \{(1, 2), (2, 3), (2, 5), (2, 6), (3, 4), (4, 6), (5, 6), (5, 7), (5, 8)\}$

Incidence, degrés, voisinage, adjacence

Une arête $e \in E$ est définie par une paire non ordonnée de sommets appelés les extrémités de e. Si l'arête e relie les sommets i et j, on dira que ses sommets sont *adjacents*, ou *incidents* avec e ou bien que l'arête e est *incidente* avec les sommets i et j. On notera alors $(i, j), \forall i, j \in V \times V$. Le nombre d'arêtes incidentes à un sommet v est appelé *degré* de v, noté $d(v)$. La matrice des *degrés* $\boldsymbol{D}$ est une matrice diagonale de taille $N_v \times N_v$ où le (i, i)-ème élément est $d_i = \sum_{j \neq i} A_{ij}$ correspond au nombre de connexions du sommet i.

Les sommets avec lesquels v partage une arête sont appelés les *voisins* de v et sont notés $\Gamma(v)$. On peut représenter un graphe G sous la forme d'une matrice d'adjacence A où chaque $A_{i,j}$ vaut 1 ou 0 s'il existe ou non une arête entre les sommets i et j. La matrice d'adjacence $\boldsymbol{A}$ du graphe G est une matrice symétrique de taille $N_v \times N_v$ dont les entrées pour le graphe d'exemple Figure 2.1 sont :

$$A_{ij} = \begin{cases} 1, & \text{si } (i,j) \in E \\ 0, & \text{si } (i,j) \notin E \end{cases} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Orientation, pondération, complétude

Les arêtes peuvent avoir un sens, dans ce cas on les appelle des *arcs* et le graphe G est dit orienté. On pourra alors distinguer le sens $u \to v$ et $v \to u$. Étant donné un arc (u, v) on dit que u est l'*origine* (ou la *source*) de (u, v) et que v est l'*extrémité* (ou la *cible*) de (u, v).

Le *demi-degré extérieur* (ou *degré sortant*) d'un nœud, noté $d^+(u)$, est le nombre d'arcs ayant le nœud u pour origine. Et inversement le *demi-degré intérieur* (ou *degré entrant*) d'un nœud, noté $d^-(u)$, est le nombre d'arcs ayant le nœud u pour extrémité. Chaque arc ayant une seule origine et une seule extrémité, le graphe comporte autant de degrés sortants que de degrés entrants : $\sum_{u \in V} d^+(u) = \sum_{u \in V} d^-(u)$.

Le *degré maximal* d'un graphe G, noté $\Delta(G)$ et le *degré minimal* de ce graphe, noté $\delta(G)$, sont respectivement le maximum et le minimum des degrés de ses sommets. Dans un graphe régulier, tous les sommets ont le même degré, et on peut donc parler du degré du graphe.

Une arête (u, u) joignant un sommet à lui-même est appelée *boucle*.

On peut également affecter un poids sur les arêtes $w(i, j)$ entier ou réel, positif ou négatif. On dit alors que le graphe est *pondéré*.

On peut alors caractériser les graphes avec ces critères. On dira d'un graphe G qu'il est *simple* s'il est non-orienté, non-pondéré et sans boucle. Il sera *complet* si tous ses sommets sont adjacents deux à deux : $\forall (u, v) \in V \times V, u \neq v \implies (u, v) \in E$.

Sous-graphe

Un graphe $H = (V_H, E_H)$ est un *sous-graphe* de $G = (V, E)$ si et seulement si $V_H \subseteq V$ et $E_H \subseteq E$. G est alors appelé le *super-graphe* de H. H est dit *induit* par V_H si $\forall u, v \in V_H, (u, v) \in E \iff (u, v) \in E_H$. Autrement dit on définit H uniquement par l'ensemble de ses sommets V_H et il contient implicitement toutes les arêtes du super-graphe G incidentes à deux extrémités de ces sommets.

2.2.2 Parcours de graphe

Définition Dans un graphe non orienté, une *chaîne* de longueur l reliant u à v, notée $\mu(u, v)$, est définie par une suite de l arêtes successivement adjacentes de E, $\mu(u, v) = \{e_i, i = 0, ..., l-1\}$ avec $\forall s \in V, e_0 = (u, s_1), e_1 = (s_1, s_2), ..., e_{l-2} = (s_{l-2}, s_{s-1}), e_{l-1} = (s_{l-1}, v)$. La notion correspondante dans les graphes orientés est celle de *chemin*.

Une *chaîne élémentaire* est une chaîne ne passant pas deux fois par un même sommet, c'est-à-dire dont tous les sommets sont distincts.

Une *chaîne simple* est une chaîne ne passant pas deux fois par une même arête, c'est-à-dire dont toutes les arêtes sont distinctes.

Dans le cas des graphes non pondérés, *longueur* et *poids* d'une chaîne sont deux notions identiques car on attribue à toutes les arêtes le même poids : 1. Cependant dans le cas d'un graphe pondéré, les poids des arêtes peuvent différer, on veillera donc à distinguer une chaîne de poids minimal d'une chaîne de longueur minimale.

Définition Un *cycle* est une chaîne simple dont les deux extrémités sont identiques. Dans le cas d'un graphe orienté, un cycle orienté est un *circuit*.

Un graphe orienté (resp. non orienté) ne possédant aucun circuit (resp. cycle) est considéré comme *acyclique*.

Dans le graphe présenté Figure 2.1 la séquence $\{(1, 2),\ (2, 6),\ (6, 5),\ (5, 8)\}$ forme une chaîne de longueur 4. La séquence $\{(2, 5),\ (5, 6),\ (6, 4),\ (4, 3),\ (3, 2)\}$ est un cycle de longueur 5.

2.2.3 Connexité

Définition Un graphe non orienté $G = (V, E)$ est dit *connexe* si $\forall u, v \in V$, il existe une chaîne reliant u à v. Une *composante fortement connexe* (CFC) de G est un sous-graphe H induit de G tel qu'il existe une chaîne reliant tout couple $(u, v), u \neq v$ de sommets H.

Pour un graphe orienté, on parle de connexité si en oubliant l'orientation des arêtes, le graphe est connexe. On parle de *forte connexité* s'il existe un chemin orienté depuis tout nœud u vers tout nœud v.

Pour mesurer la "force" de la connexité d'un graphe non orienté, on s'intéresse au nombre de sommets (d'arêtes) dont la suppression est nécessaire pour que le graphe ne soit plus connexe. On appelle sommet-connexité (parfois connectivité) le nombre minimum de sommets dont l'élimination déconnecte G ou le réduit à un sommet unique. On note la connectivité $\kappa(G)$. Ainsi un graphe connexe G sera dit *k-sommet-connexe* ($k > 0$), si $\kappa(G) \geq k$. Un graphe connexe G sera dit *k-arête-connexe* s'il ne peut être déconnecté par l'élimination de moins de k arêtes. Un graphe *k-sommet-connexe* est *k-arête-connexe*. La Figure 2.2 représente un graphe contenant deux composantes fortement connexes. Le sous-graphe engendré par l'ensemble de sommets $\{1, 2, 3, 4, 5\}$ est 1-sommet-connexe et par $\{6, 7, 8\}$ est 2-sommet-connexe.

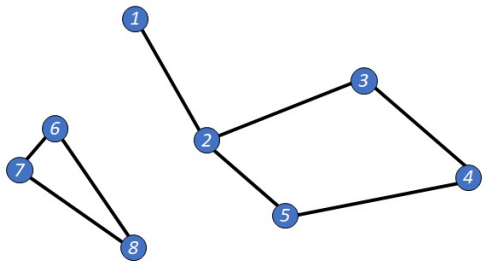


FIGURE 2.2 – Un graphe non connexe, contenant deux composantes fortement connexes

2.2.4 Matrices et réseaux de graphes

Il est utile dans la modélisation et l'analyse de données de réseau de pouvoir caractériser un graphe et certains aspects de sa structure à l'aide de matrices. Nous présentons ici deux matrices fondamentales en théorie des graphes, la matrice d'adjacence $\boldsymbol{A}$ et la matrice Laplacienne $\boldsymbol{L}$, qui permettent de caractériser et d'analyser la structure d'un graphe.

La matrice d'adjacence $\boldsymbol{A}$ introduite précédemment Section 2.2.1, permet de capturer la connectivité fondamentale d'un graphe G dans une matrice symétrique de taille $N_v \times N_v$ dont les entrées sont :

$$A_{ij} = \begin{cases} w(i,j), & \text{si } (i,j) \in E \\ 0, & \text{si } (i,j) \notin E. \end{cases}$$

Lorsque le graphe G est non pondéré $w(i, j) = 1$. Dans le cas d'un graphe pondéré, $w(i, j) \in \mathbb{R}$ représente le poids qui sépare les noeuds i et j par exemple la longueur d'une canalisation.

La matrice des *degrés* $\boldsymbol{D}$ est une matrice diagonale de taille $N_v \times N_v$ où le (i, i)-ème élément est $d_i = \sum_{j \neq i} A_{ij}$ correspond au nombre de connexions du sommet i.

Soit $\boldsymbol{A}^r$ la r-ème puissance de $\boldsymbol{A}$, alors les entrées $\boldsymbol{A}^r_{ij}$ représentent le nombre de chaînes possibles de taille r entre les noeuds i et j. On peut aussi définir une matrice d'adjacence dans le cas d'un graphe dirigé. Dans ce cas $A_{ij} = 1$ s'il existe une arête dirigée de i vers j. A_{ij} n'étant plus égale à A_{ji}, $\boldsymbol{A}$ n'est plus symétrique.

La matrice Laplacienne d'un graphe s'exprime de la forme :

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}.$$

$\boldsymbol{L}$ est symétrique car $\boldsymbol{D}$ et $\boldsymbol{A}$ sont deux matrices symétriques.

De plus pour tout vecteur $u \in R^{N_v}$, on a :

$$\begin{aligned} u^T L u &= u^T D u - u^T A u \\ &= \sum_{i=1}^{N_v} d_i u_i^2 - \sum_{\{i,j\}\in E} u_i A_{i,j} u_j \\ &= \frac{1}{2}\Big(\sum_{i=1}^{N_v} d_i u_i^2 - 2\sum_{\{i,j\}\in E} u_i u_j A_{i,j} + \sum_{j=1}^{N_v} d_j u_j^2\Big) \\ &= \frac{1}{2}\sum_{\{i,j\}\in E} A_{i,j}(u_i - u_j)^2 \geq 0 \end{aligned} \tag{2.2.1}$$

La matrice Laplacienne est alors semi-définie positive et diagonalisable. Toutes les valeurs propres de $\boldsymbol{L}$ sont positives ou nulles et on note $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{N_v}$ les valeurs propres ordonnées de $\boldsymbol{L}$.

$\boldsymbol{L}$ admet la diagonalisation $L = \Phi \Delta_L \Phi^T = \sum_{i=1}^{N_v} \lambda_i \phi_i \phi_i^T$ où la matrice diagonale contient les valeurs propres $\lambda_i, i = 1 \ldots N_v$ et $\Phi = [\phi_1|, \ldots, |\phi_{N_v}]$ est la matrice orthogonale des vecteurs propres normés : $\Phi^T \Phi = Id_{\mathbb{R}^{N_v}}$.

Le vecteur propre $\phi_1 = (u_1, \ldots, u_{N_v})$ associé à la seule valeur propre nulle est constant : $\phi_1 = 1/\sqrt{N_v}\mathbf{1}$. ϕ_i est un vecteur propre de $\boldsymbol{L}$. En effet si ϕ_1 est un vecteur propre de L associé à la valeur propre 0, on a $L\phi_1 = 0$ et d'après (2.2.1) :

$$\phi_1^T L \phi_1 = 0 = \sum_{\{i,j\}\in E} A_{i,j}(u_i - u_j)^2$$

Puisque $A_{i,j} \geq 0$, la somme est nulle si tous ses termes sont nuls, c'est à dire si et seulement si on a $A_{i,j}(u_i, u_j)^2 = 0, \forall i, j \in V$. Si l'arête $\{v_i, v_j\} \in E$ alors le poids $A_{i,j}$ est non nul et nécessairement $u_i = u_j$. Donc le vecteur est propre $\phi_1 \in \mathbb{R}^{N_v}$ est constant sur les coordonnées correspondant à des nœuds connectés dans le graphe.

Les vecteurs propres forment une base orthonormale $\phi_i^t \phi_j = \gamma_{ij}$ pour tout vecteur propre on a $\phi_i^t \mathbf{1}_{N_v} = 0, 2 \leq i \leq N_v$

par conséquent les vecteurs propres $\phi_i, 2 \leq i \leq N_v$ vérifient : $\sum_{j=1}^{N_v} \phi_i(u_j) = 0$

Les propriétés spectrales de $\boldsymbol{L}$ de ses vecteurs et valeurs propres fournissent des informations sur la structure d'un graphe G que l'on présente Section 2.4.2.

2.3 Du réseau d'eau potable au réseau de graphe

Un grande partie des travaux de thèse se sont concentrés sur l'élaboration d'une structure de données unique regroupant l'ensemble données présentées en Section 1.4. L'idée étant d'avoir à disposition une structure unique sur laquelle représenter les données souhaitées de façon simple et interprétable et permettant l'utilisation et le croisement d'un maximum de données issues de différentes bases de données à l'aide de la théorie des graphes.

La structure de données de graphe possède de multiples avantages quand il s'agit de prendre en compte les données générées autour de la gestion d'un réseau d'eau potable. Un des plus immédiats est que la structure composée de nœuds et d'arêtes permet de transposer directement le squelette du réseau, composé d'accessoires et de canalisations dans la base de données. Les liens entre tous les objets du réseau sont alors intrinsèquement encodés par cette structure. Si deux canalisations sont connectées par le biais d'un accessoire, dans la bases de données deux arêtes seront inter-connectées par un nœuds.

De ce fait il est alors possible de représenter toutes les données structurelles du réseau AEP sur la structure du graphe. Comme il est présenté section 2.3.2 et 2.3.3, les nœuds et arêtes du graphes sont labellisés lors de leur création par les données issues de différentes bases de données. Le réseau ainsi créé permet une représentativité immédiate du squelette du réseau, par exemple sous la forme d'un graphique nœud arête géo-localisable (voir Figure 2.5).

Comme présenté dans le chapitre précédent, beaucoup de données sont générées autours de la gestion du réseau AEP. Il peut s'agir de chroniques temporelles issues de mesures faites par des capteurs ou d'informations liées aux volumes consommés par un client. Ces données sont récoltées, générées ou utilisées par différents services en fonction de leurs missions. Chaque service dispose de ses outils, il n'existe pas de format spécifique pour ces données et il existe une inégalité dans l'exactitude et la cohérence de celles-ci vis à vis de la structure du réseau. Par exemple le but du SIG est de représenter le plus fidèlement et à jour possible le patrimoine réseau à son délégataire, sans forcément prendre en compte toutes les informations relatives aux modifications court terme du réseau. Le service clientèle s'intéresse à la facturation du volume consommé, peu importe le type de compteur ou à quel branchement il est raccordé. Le service qualité eau va s'intéresser aux chroniques temporelles mesurées par les capteurs pour identifier et anticiper les risques.

Pourtant toutes ces données ne sont pas connectées entre elles, et sont parfois difficilement localisables sur le réseau AEP. Une grande partie des travaux de thèse s'est concentrée sur l'identification, la collecte et le traitement de ces données afin de les intégrer dans une structure unique.

2.3.1 Structure des données de graphe

Il existe deux structures de données communes pour représenter un graphe G. La première est la matrice d'adjacence $\mathbf{A}$ de taille $N_v \times N_v$ définie précédemment section 2.2.4. Ce choix est souvent naturel, étant donné que les matrices sont des objets fondamentaux dans la plupart des environnements de programmation et de logiciels.

Toutefois, si le graphe est volumineux, et particulièrement si celui-ci est creux (i.e., si $N_e \sim N_v$), il peut être préférable d'utiliser une collection de listes d'adjacence. En effet,

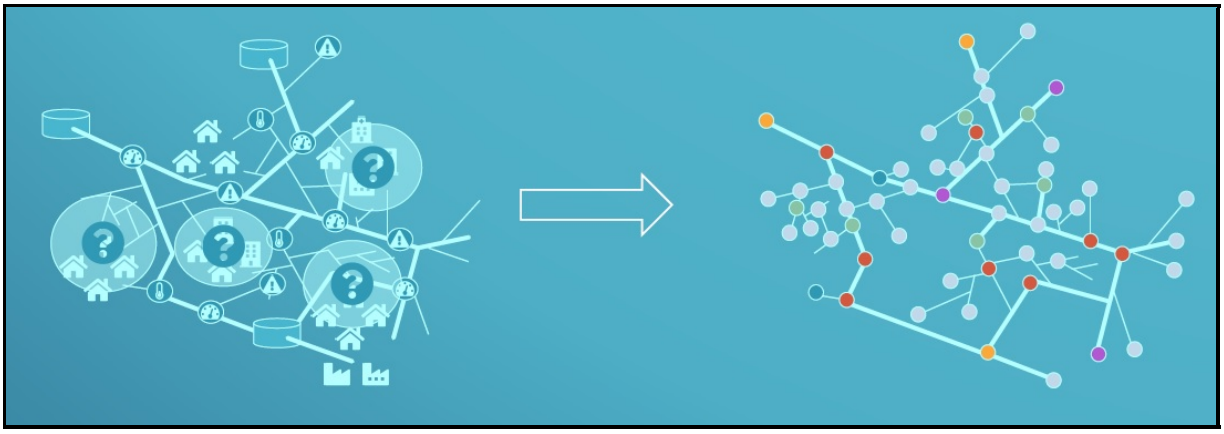


FIGURE 2.3 – Du réseau d'eau potable au réseau de graphe

dans ce cas, la matrice d'adjacence sera à la fois volumineuse et remplie principalement de zéros, car elle représente explicitement les arêtes présentes et absentes, tandis que les listes d'adjacence stockent uniquement les informations sur les arêtes présentes.

A la vue de la topologie spécifique des réseaux AEP, où toutes les canalisations sont connectées entre elles par le biais d'accessoires, on se retrouve alors avec une matrice d'adjacence qui sera creuse. Au maximum un raccord en croix ne raccordera que quatre canalisations entre elles $\Delta(G) = 4$.

Il existe une multitude d'outils, bibliothèques ou API permettant à partir de liste de noeuds ou d'arêtes de représenter et analyser des réseaux de graphe. Durant les travaux de thèse nous avons pu utiliser trois API différentes permettant chacune de répondre à des besoins spécifiques.

Nous avons commencé par utiliser *IGraph* une collection de bibliothèques permettant de créer et manipuler des réseaux de graphe. Cette bibliothèque est utilisée en tant que package sur le logiciel de développement statistiques **R**. Cette première étape nous a permis d'effectuer à la fois la collecte et le traitement de l'ensemble des données brutes présentées précédemment. Néanmoins cette bibliothèque possède quelques limites concernant la visualisation de grands graphes.

Concernant la visualisation des données nous avons utilisé deux API développées par le LaBRI, Tulip et Wulip. Le graphe *Bx* comportant plus de 60 000 nœuds il n'était pas possible de le visualiser avec *IGraph*.

2.3.2 Les arêtes du graphe

Nous allons dans cette partie formaliser la création de l'ensemble des arêtes E du graphe G. Pour chaque canalisation présente dans le SIG nous allons créer une arête. Cette arête possède alors deux extrémités qui seront détaillées dans la section suivante. Le réseau *Bx* est ainsi composé de $N_v(Bx) = 64783$ arêtes et celui d'*Ambès* de $N_v(Ambès) = 1188$.

Afin d'avoir une représentation exacte de la structure du réseau AEP sous la forme d'un graphe, sa structure devrait être représentée jusqu'aux compteurs. Pour cela il serait nécessaire de rajouter autant d'arêtes que le réseau AEP compte de branchements. Mais cela comporte deux problématiques.

La première est qu'il existe un très grand nombre de branchements sur le réseau AEP,

presque trois fois plus que le nombre de canalisations. Comme indiqué dans la section 1.4.1, le réseau AEP de la métropole bordelaise est composé de 188 200 branchements. Ce traduisant par l'ajout d'autant d'arêtes rendant ainsi le graphe très large.

La seconde problématique est que la position exacte du raccordement entre une canalisation et chacun de ses branchements n'est pas connue avec précision. Il n'existe pas d'objet physique dans le SIG représentant la connexion entre une canalisation et un branchement, comme cela peut être le cas entre deux canalisations par le biais d'un raccord.

Il faudrait alors rajouter autant de noeuds qu'il existe de branchements. Une arête (i, j) étant raccordée à p branchements se verra alors divisée en $p - 1$ arêtes, représentant des segments canalisations virtuels. Avec une moyenne de 5 branchements par canalisation et un maximum de 84 branchements sur une canalisation du réseau *Bx*, cette transformation rendrait la structure du graphe inutilement plus complexe.

Il est alors plus pertinent de regrouper l'ensemble des branchements aux canalisations auxquelles elles sont raccordées. Ainsi toutes les informations des branchements sont agrégées sur un nœud. Ce nœud "virtuel" permet de représenter l'existence d'un ou plusieurs branchements raccordés sur la canalisation. Par souci de simplification les nœuds virtuels sont placés au centre de chacune des canalisations concernées.

Ainsi si une arête (i, j) possède au moins un branchement, deux arêtes adjacentes (i, n) et (n, j) sont alors crées et donc un noeud virtuel n regroupant les branchements sur cette canalisation comme l'illustre la Figure 2.4. Les deux arêtes possèdent l'ensemble des caractéristiques de l'arête initiale (diamètre, date de pose, matériau etc.) mais auront une longueur divisée par deux. Sur les 64 783 arêtes du graphes 35 307 possèdent au moins un branchement raccordé, avec en moyenne 5.4 branchements par arête. Finalement 35 307 nœuds virtuels sont rajoutées au graphe représentant le réseau AEP de la métropole bordelaise (voir Table 2.1).

Graphe	Canalisation	Branchement	Canalisation avec branchements	Branchement par canalisation		Arête
				moyen	maximum	
Bx	63 595	186 660	34 586	5.4	84	98 191
Ambès	1 188	2 364	721	3.3	23	1 188
Total	64 783	189 024	35 307	5.4	84	99 379

TABLE 2.1 – Statistiques du réseau de graphe de BM

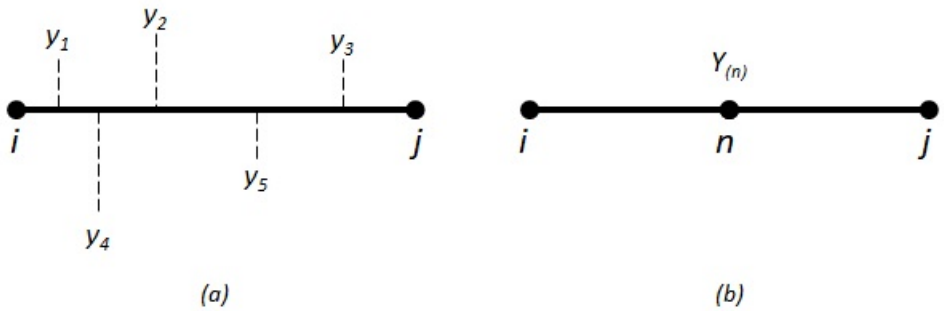


FIGURE 2.4 – Illustration de l'agrégation des branchements au niveau du nœud virtuel

Finalement nous disposons pour chacune des arêtes d'un identifiant unique : l'identi-

fiant issu du SIG permettant de croiser les informations entre le graphe et l'ensemble des données du SIG.

Bien que nous avons maintenant une structure de graphe créée à partir de la liste des arêtes il est nécessaire de créer la liste des nœuds du graphe. Celle-ci permettant la labélisation des nœuds du graphe à l'aide d'informations relatives au type de nœuds considérés. Si l'objectif était seulement d'étudier la topologie du réseau AEP par le biais de ses canalisations nous serions en mesure de l'effectuer à l'aide de cette liste d'arêtes.

ID.SIG	Extr1	Extr1.type	Extr2	Extr2.type
7/10	8	RACCOR	10	RACCOR
29965/9	8	VANNE	9	RACCOR
29966/9	9	VANNE	9	RACCOR
9/17	9	RACCOR	16	RACCOR
11/29979	21	VANNE	11	RACCOR

TABLE 2.2 – Résumé de la table représentant la liste des arêtes, avec leurs identifiant dans le SIG ainsi que ceux des accessoires présents à chaque extrémité avec leur type.

2.3.3 Les nœuds du graphe

Nous allons partir de la liste des arêtes précédemment créée afin d'obtenir la liste complète des nœuds du graphe. Le SIG étant sous la forme d'une base de données structurée pour chaque accessoire du réseau AEP nous disposons d'une table relationnelle indiquant quelles tables et quels éléments sont connectés entre eux (cf. Table 2.2). On est donc en mesure de déterminer pour chaque canalisation quel type de nœud est présent à ses extrémités. Pour chaque canalisation peut exister une vingtaine de tables de données en relation. Nous allons ici nous concentrer sur les tables représentant les relations physiques avec les canalisations à savoir, les raccords, les vannes, les appareils de mesures, les réservoirs et les usines de production d'eau potable.

La 1ère étape consiste à parcourir la liste des arêtes et pour chaque objet à ses extrémités à créer un noeud. Par exemple, la canalisation avec l'identifiant ID.SIG 29965/9 ligne 2 de la Table 2.2 possède à ses extrémités une vanne et un raccord ayant pour identifiant 8 dans la table des vannes et respectivement 9 dans la table des raccords.

La 2ème étape consiste à étiqueter les différents nœuds du graphe. A l'aide des identifiants on récupère leurs coordonnées nous permettant de les placer sur un plan en deux dimensions mais aussi un certain nombre variable nous permettant de labéliser l'ensemble des noeuds en fonction de leurs types. Si c'est une vanne son état : ouvert/fermé ; un capteur : l'ensemble des paramètres mesurés sous la forme de séries chronologiques ; un réservoir : la quantité d'eau stockée au court du temps etc..

La 3ème étape consiste à ajouter et labéliser les noeuds virtuels représentant les points de consommation d'eau raccordés à chacune des canalisations. Comme indiqué dans la Section 1.4.2.b sur les volumes sortants, il est nécessaire de faire un croisement entre la base de données clientèle et SIG afin de connaître le plus précisément possible le nombre de compteurs en fonctionnement branchés sur chacune des canalisations. On différenciera alors deux variables pour chacun des noeuds sortants, une représentant les volumes mesurés à l'aide de compteurs équipés de la télé-relève et une autre représentant les volumes

consommés obtenus à partir de modèles d'estimation de consommations. On est donc capable pour chacun des noeuds virtuels de définir le volume total sortant représentant les volumes mesurés par l'ensemble des compteurs. Les noeuds virtuels sont tous placés au centre des arêtes auxquels elles sont raccordées et sont tous de degré $\delta(i) = 2$.

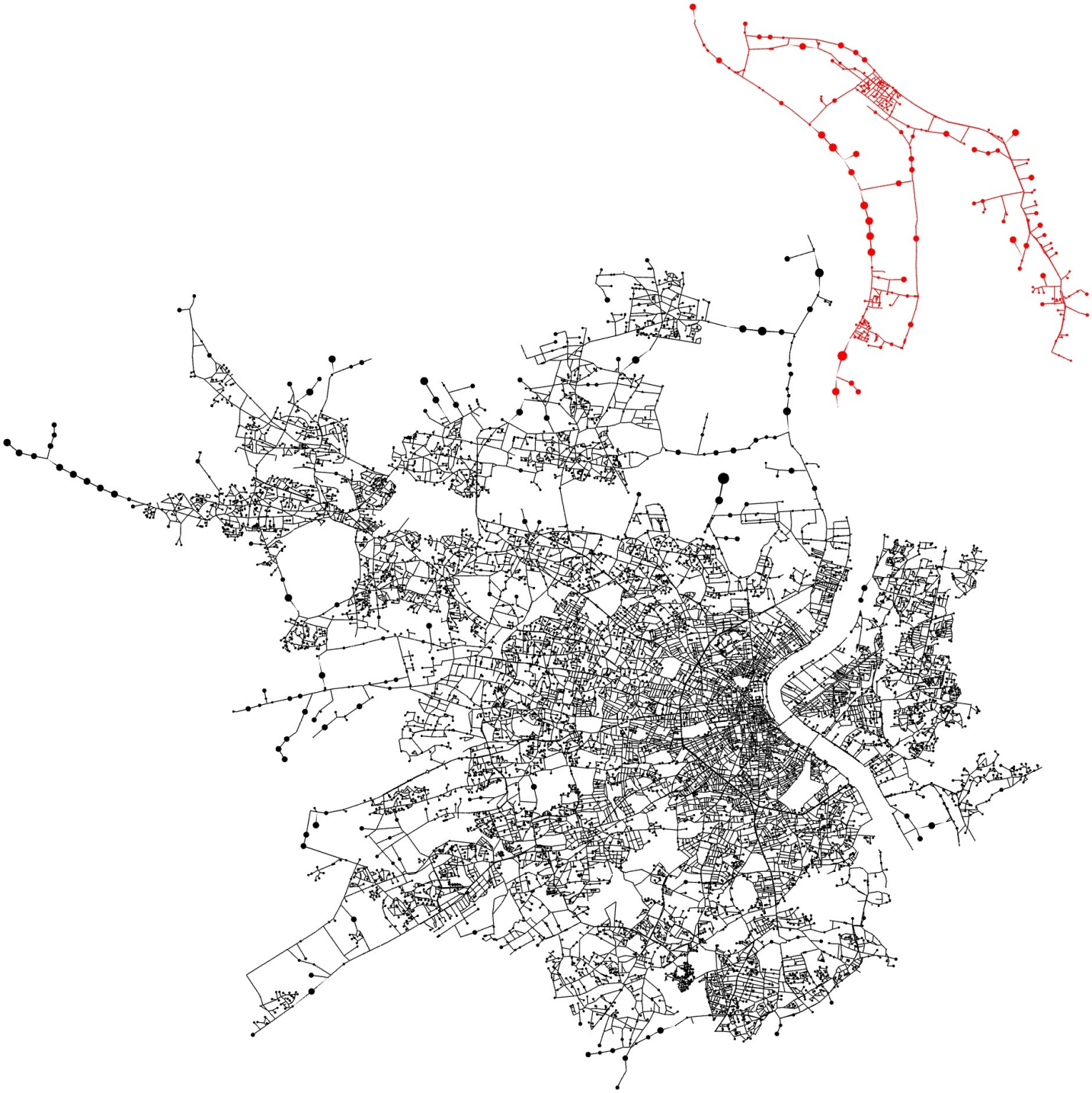


FIGURE 2.5 – Représentation sous la forme sommet et arête des réseaux AEP *Bx* en noir, et *Ambès* en rouge. Chaque arête représente une canalisation du réseau et chaque nœud un objet du réseau connecté à une ou plusieurs canalisations. (Visualisation géo-spatiale des nœuds issue du logiciel Tulip)

2.4 Analyse descriptive du réseau de graphe

Le réseau de graphe utilisé durant la thèse fournit une représentation de l'état structurel du réseau tel qu'il était en décembre 2017. Cette date a été sélectionnée car il s'agissait de la dernière mise à jour des informations structurelles du réseau dans la basse de données SIG.

Il existe de multiples méthodes pour dessiner des graphes en utilisant des symboles pour les noeuds $v \in V$ et des courbes pour les arêtes $e \in E$. La Figure 2.5 fournit une représentation graphique sous la forme sommets et arêtes des deux réseaux étudiés (en rouge d'*Ambès* et en noir *Bx*). Les nœuds y sont positionnés par rapport à leur position géo-spatiale tel qu'elle est indiquée dans la base de données SIG et les arêtes par des droites reliant les nœuds connectés entre eux.

De nombreux travaux explorent une variété de stratégies pour mieux concevoir, comprendre la robustesse et les propriétés structurelles des réseaux AEP, en particulier en ce qui concerne une éventuelle défaillance, à l'aide de la théorie des graphes [Hawick, 2012, Dunn and Wilkinson, 2013, Yazdani and Jeffrey, 2012a].

Dans l'ensemble, ces travaux se basent sur l'analyse topologique des réseaux pour comprendre et analyser la structure des systèmes de distribution d'eau potable.

[Yazdani and Jeffrey, 2011] proposent l'utilisation des mesures structurelles issues de la théorie des graphes pour quantifier des propriétés telles que la redondance et la connectivité optimale pour l'optimisation de la conception d'un réseau AEP.

[Fortini et al., 2014] utilisent une méthode de squelettisation du réseau réduisant considérablement sa complexité. Le réseau simplifié est ensuite analysé pour comprendre ses principaux modèles de flux ou comme première étape pour déterminer ses éléments les plus vulnérables ou les plus importants.

[Herrera et al., 2016] définissent un indicateur de résilience pour les réseaux AEP sectorisés basé sur la quantification de la redondance et de la capacité de toutes les routes possibles depuis les nœuds de demande jusqu'à leurs sources d'approvisionnement. La mesure quantitative obtenue permet de déterminer la qualité de la connexion d'un nœud aux sources d'eau disponibles sur le réseau.

Bien que nos travaux de thèse ne se soient pas focalisés sur l'analyse topologique des réseaux étudiés nous avons utilisé la structure de graphe et la représentation de données spécifiques sur les nœuds et arêtes de celui-ci pour répondre à des problématiques métiers. Nous nous proposons dans cette section d'explorer quelques caractéristiques et propriétés structurelles des réseaux *Ambès* et *Bx*, ainsi que de présenter les avantages que l'on peut tirer d'une telle structure pour des données issues d'un réseau AEP. Nous présentons aussi deux cas d'application d'utilisation de graphe complexe pour la création de chantiers opérationnels de renouvellement et l'ajustement d'estimations de consommation d'eau potable.

2.4.1 Propriétés structurelles du graphe

Les nœuds et arêtes d'un graphe constituent les éléments fondamentaux, il existe un certain nombre de mesures que l'on peut obtenir d'eux nous permettant de caractériser un graphe.

Degré Le degré d_v d'un nœud v, dans un graphe $G = (V, E)$, compte le nombre d'arêtes dans E incidentes à v. Cette mesure fournit une quantification basique de la façon dont un nœud v est connecté aux autres nœuds du graphe. On peut alors considérer la séquence de degré du graphe $\{d_1, \ldots, d_{N_v}\}$ et obtenir des informations sur la connectivité globale du graphe.

La distribution des degrés fournit un résumé naturel de la connectivité d'un graphe. La Figure 2.6 représente la distribution des degrés des graphes *Ambès* en orange et *Bx* en bleu.

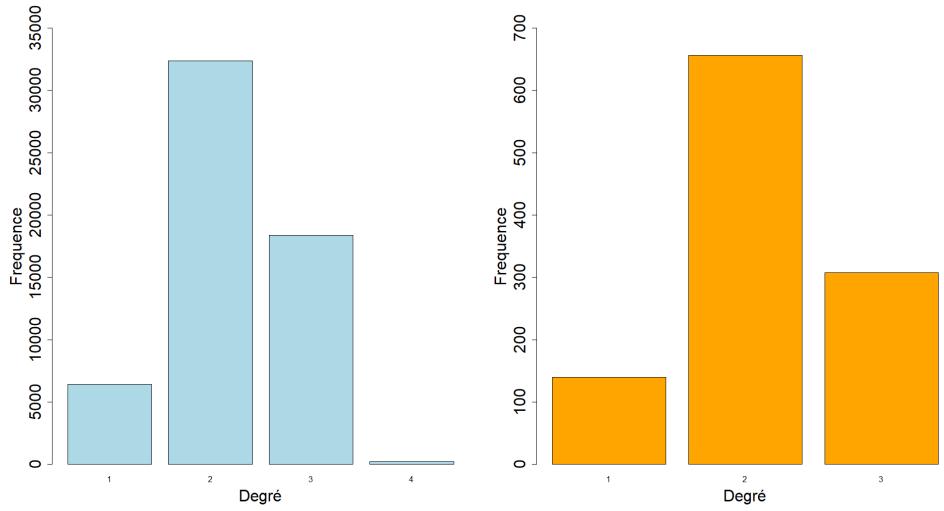


FIGURE 2.6 – Distribution des degrés pour les graphes *Bx* en bleu et *Ambès* en orange

Les deux réseaux bien que de volumétries différentes $N_{v_{Bx}} > N_{v_{Amb}}$ sont d'une connectivité relativement similaire et élémentaire. Bien que le graphe *Ambès* ne dispose pas de nœuds de degrés $d = 4$, ils ont tout deux une répartition des degrés relativement similaire. Les deux réseaux sont composés d'une grande proportion de nœuds de degrés $d = 2$ représentant les canalisations connectées deux à deux. Les nœuds de degrés $d = 3$ et $d = 4$ illustrent la structure maillée des réseau AEP, et les nœuds de degrés $d = 1$ les queues du réseau.

Pour les réseaux pondérés, une généralisation du degré est la notion de force d'un nœud, qui est obtenue simplement en additionnant les poids $w_{i,j}$ des arêtes incidentes à un sommet donné. La répartition de la force, aussi appelée degré pondéré, est définie par analogie avec la distribution des degrés. La force des nœuds pour les graphes *Bx* et *Ambès* est représentée sur la figure 2.7 à gauche.

Ici la longueur des canalisations est utilisée comme pondération, $w_{i,j} = l_{i,j}$. On observe pour les deux réseaux des degrés pondérés similaires avec une tendance pour les arêtes de faible longueur à être connectées entre elles. Cela s'explique par le fait que le réseau est composé en grande partie de canalisations de moins de 100m comme on a pu l'illustrer figure 1.6.

Au-delà de la distribution de degrés il peut être intéressant de comprendre la manière dans laquelle des sommets de différents degrés sont liés les uns aux autres. Afin d'évaluer cette caractéristique nous pouvons utiliser la notion de degré moyen des plus proches voisins d'un sommet donné (Average Nearest Neighbor Degree, ANND en anglais, [Newman,

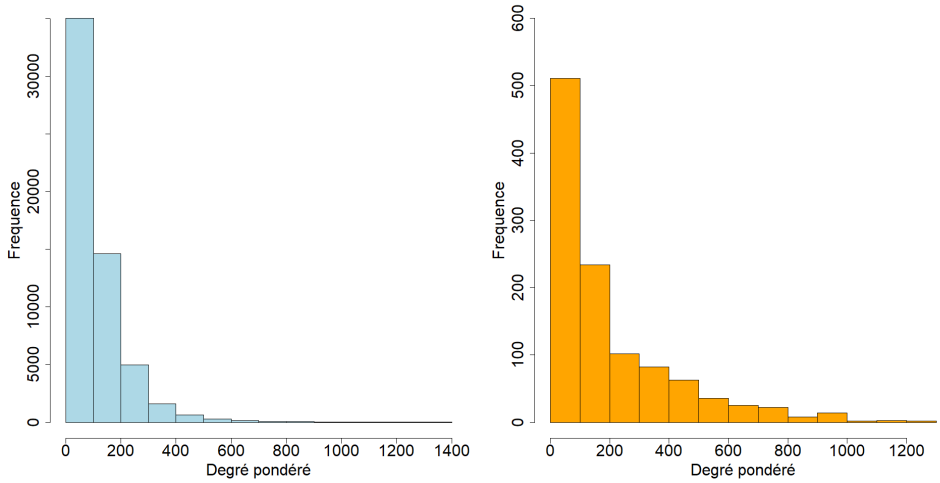


FIGURE 2.7 – Distribution des degrés pondérés avec la longueur des arêtes en mètre, pour les graphes *Bx* en bleu et *Ambès* en orange

2002]). La figure 2.8 représente la répartition des ANND par rapport à la répartition des degrés du graphe.

Pour les deux réseaux on constate que $\approx 65\%$ des nœuds de degré $d = 1$ sont connectés à des nœuds de degrés 2. Les nœuds de degrés $d = 2$ sont pour les deux graphe dans 66% des cas connectés à des noeuds de degré $d \geq 2$. Les nœuds de degrés $d = 3$ et $d = 4$ ont quant à eux tendance à être connectés à des nœuds de plus faible degrés, principalement des nœuds de degrés $d = 2$.

Distance On peut aussi caractériser un graphe à l'aide de notions de distance entre les nœuds qui le compose.

En théorie des graphes, le diamètre d'un graphe est la plus grande distance possible qui puisse exister entre deux de ses sommets. Cette distance est définie par la longueur d'un plus court chemin entre ses sommets. Pour cela on calcule l'excentricité de chacun des nœuds du graphe $\epsilon(v) = \max_{u \in V} d(v, u)$. Ainsi le diamètre $diam$ d'un graphe correspond à la plus grand excentricité du graphe $diam = \max_{v \in V} \epsilon(v)$.

Pour le graphe de *Bx* $diam_{Bx} = 730$ (resp. *Ambès* $diam_{Amb} = 220$) ce qui signifie que la plus grande distance séparant deux nœuds dans le graphe de *Bx* (resp. *Ambès*) est composée de 730 arêtes (resp. 220 arêtes). Cette notion de diamètre peut être pondérée, si on observe cette distance avec comme poids $w_{u,v} = l_{u,v}$ la longueur des canalisations on observe que la distance la plus longue entre deux noeuds du graphe est de $36Km$ pour le réseau *Bx* et $20Km$ pour celui d'*Ambès*

Le rayon d'un graphe est la plus petite des excentricités du graphe, $r = \min_{v \in V} \epsilon(v)$. Pour le graphe *Bx*, $r_{Bx} = 250$ (resp. *Ambès*, $r_{Amb} = 66$).

Ces notions de distance sont intéressantes pour évaluer l'étendue d'un graphe. Les distances calculées précédemment sont calculées dans des graphes non dirigés. C'est à dire que l'on calcule la distance séparant un nœud à n'importe quel autre nœud du graphe tant qu'ils sont connectés.

Dans le cas des réseaux AEP, les sens d'écoulement varient au court du temps, ce qui

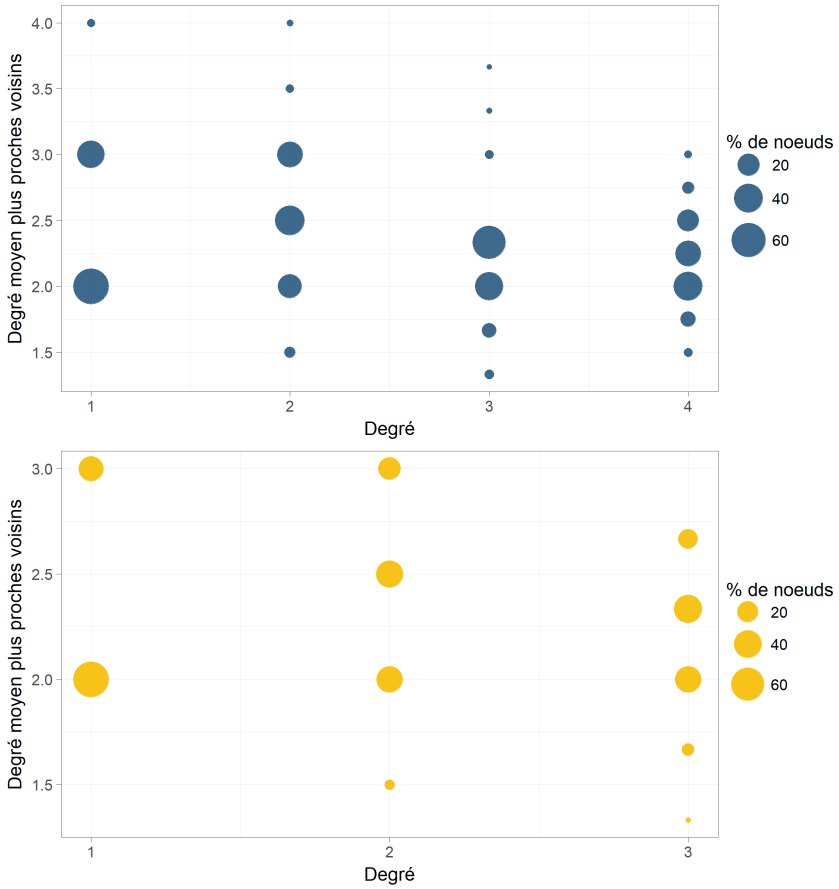


FIGURE 2.8 – Degré du graphe en fonction des degrés moyens des plus proches voisins (ANND) pour le graphe *Bx* en bleu et *Ambès* en orange. La proportion de nœud du même degré est calculé par colonne.

implique que les sens des arcs sont pas connus et mesurés à chaque instant. Cette notion d'excentricité pourrait s'appliquer à un graphe pour lequel les débits et sens d'écoulement sont connus et mesurés. On pourrait alors calculer l'excentricité dans le cas d'un graphe orienté et déterminer le temps de séjour le plus long pour chacun des nœuds du graphe.

Centralité Les indicateurs de centralité sont des mesures censées capturer la notion d'importance dans un graphe, en identifiant les sommets les plus significatifs. Les indices de centralités fournissent une fonction réelle sur les sommets d'un graphe, où les valeurs fournissent un classement qui identifie les nœuds les plus importants [Bonacich, 1987, Borgatti, 2005]. Plusieurs interprétations de l'importance ont été proposés en fonction du type de réseau étudié. L'importance peut servir à classer les sommets selon leur capacité de transfert à travers le réseau ou encore comme un degré de participation à la cohésion du réseau.

Nous présentons ici le calcul des deux indices de centralité les plus communément utilisés, la centralité intermédiaire et de proximité.

Réseau	Diamètre	Rayon	$\bar{\epsilon}$
Bx	730	164	250
Ambès	220	47	66

TABLE 2.3 – Excentricité maximale, minimale et moyenne des réseaux *Bx* et *Ambès*

Centralité intermédiaire La centralité intermediaire (Betweenness centrality en anglais) compte le nombre de fois où un nœud agit comme un point de passage le long du plus court chemin entre deux autres nœuds quelconques. La centralité d'intermédiarité d'un sommet v la plus commune, introduite par Freeman [Freeman, 1977],est donnée par l'expression :

$$C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

avec $\sigma(s,t|v)$ le nombre total de plus court chemin de s à t, passant par v et $\sigma(s,t) = \Sigma_v \sigma(s,t|v)$.

Finalement la mesure de centralité intermédiaire résume à quel point un sommet est localisé "entre" d'autres pairs de nœuds. Cette mesure de centralité peut être normalisée en divisant $C_B(v)$ par un facteur $(N_v - 1)(N_v - 2)/2$.

Le graphe de gauche Figure 2.9 présente la mesure de centralité intermédiaire normalisée $\frac{C_B}{(N_v-1)(N_v-2)/2}$ pour chacun des nœuds du graphe d'*Ambès*. La couleur des nœuds indiquant la valeur du critère normalisée entre 0 et 1. Les arêtes du graphe ont été colorés avec une interpolation de la couleur entre les deux nœuds qu'elles relient afin de mettre en avant les chemins passant par les nœuds ayant le critère le plus élevé.

Cette mesure nous permet de voir, sans prendre en compte de notion de flux hydraulique, quels sont les chemins les plus susceptibles d'être empruntés par un flot circulant sur ce réseau. Les flux hydrauliques ne dépendant pas seulement de la structure du réseau mais aussi des sources d'eau dans celui-ci on ne peut comparer de façon immédiate le critère à ces flux. Néanmoins les chemins mis en lumière par le critère de centralité intermédiaire (dont la couleur tire vers le rouge) font partie des canalisations acheminant les flux d'eau sur les différents secteurs de consommation. Il s'agit donc bien de canalisations importantes du point de vue criticité, perdre ses canalisations priverait d'eau beaucoup de points du réseau.

Centralité de proximité La mesure de centralité de proximité, (Closeness centrality, en anglais), tente de mesurer à quel point un nœud est "proche" des autres nœuds du graphe.

L'approche standard est introduite par Sabidussi [Sabidussi, 1966]. Elle consiste à mesurer le chemin le plus court entre tous les nœuds d'un graphe, puis d'assigner un score basé sur la somme des plus courts chemins. On peut donc calculer pour chacun des noeuds du graphe le critère de proximité C_{Cl} :

$$C_{Cl}(v) = \frac{1}{\sum_{u \in V} dist(v,u)}$$

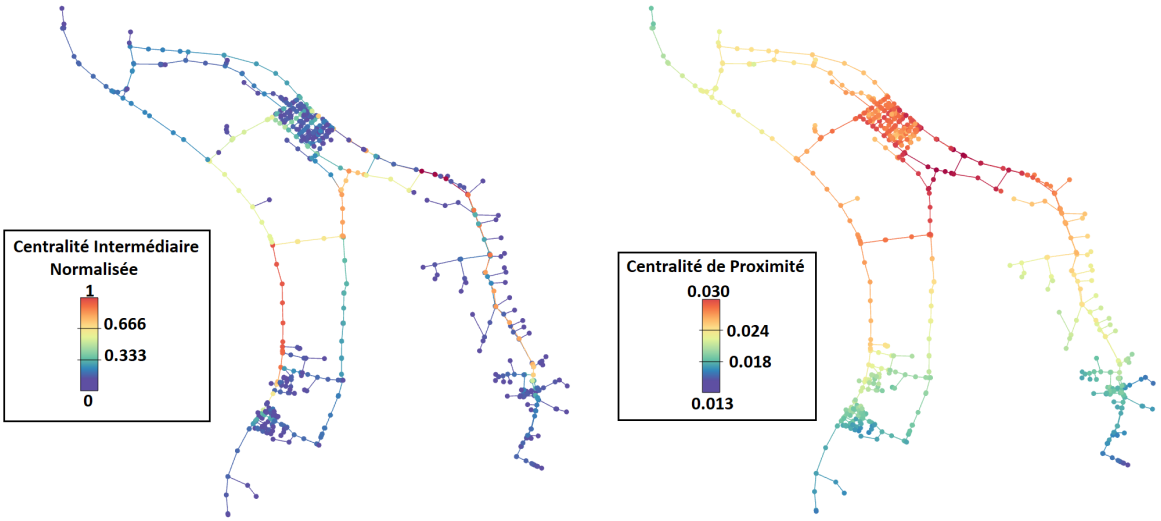


FIGURE 2.9 – Centralité intermédiaire normalisée et centralité de proximité pour le réseau *Ambès*

Où $dist(v, u)$ est la distance entre les noeuds $u, v \in V$. Plus la distance qui sépare un nœud de tous les autres nœuds du graphe est élevée plus la mesure de proximité est faible.

La mesure de centralité de proximité obtenue pour le graphe d'*Ambès* est présentée à droite Figure 2.9. Cette mesure semble relativement triviale pour une graphe représentant un réseau AEP, en effet les nœuds les plus éloignés des autres étant ceux positionnés en bout de réseau. Elle fournit tout de même une information importante sur les nœuds critiques ou à surveiller du réseau. Les nœuds avec un score de proximité sont les nœuds ayant un potentiel impact sur un maximum du nœud du réseau. Ce sont donc des points du réseau qu'il est important de sécuriser pour par exemple endiguer la propagation d'événement sur le réseau. Néanmoins comme pour la mesure de centralité intermédiaire celle-ci n'a pas été contrainte par les flux pouvant circuler dans le réseau et se base seulement sur sa topologie. Cette mesure met en avant les points du réseau qui peuvent avoir un impact important sur leurs voisinages de part leur connectivité.

Les calculs de mesures de centralité, présentés ci-dessus, requièrent le calcul des plus courts chemins entre tous les nœuds du graphe, nécessitant $\mathcal{O}({N_v}^3)$ en temps et $\mathcal{O}({N_v}^2)$ en espace. Ces calculs deviennent trop couteux en temps et mémoire pour des grands graphes comme celui de *Bx*. [Brandes, 2001] propose un algorithme permettant l'obtention des mesures de centralité qui requiert $\mathcal{O}(N_v + N_e)$ en espace et s'exécute en $\mathcal{O}(N_v N_e)$ et $\mathcal{O}(N_v N_e + {N_v}^2 log N_v)$ étapes pour un graphe respectivement non pondéré et pondéré.

Les mesures de centralité présentées Figure 2.10 sont obtenues à l'aide de l'algorithme de Brandes pour le réseau *Bx*. La mesure de centralité intérmédiaire est obtenue en 316.78 secondes et 1333.05 secondes pour un graphe respectivement non pondéré et pondéré avec la longueur des canalisations. Pour la mesure de proximité en 150.47 secondes et 880.90 secondes respectivement pour le graphe pondéré et non pondéré. Ces calculs ont été effectués sur *R* avec le package *igraph* sur une machine "classique" disposant de 8 cœurs et 16Go de mémoire de vive.

Pour aller plus loin il serait intéressant d'effectuer ces calculs par secteur hydraulique afin d'obtenir des mesures de centralité locales. On pourrait aussi comme suggéré avec la

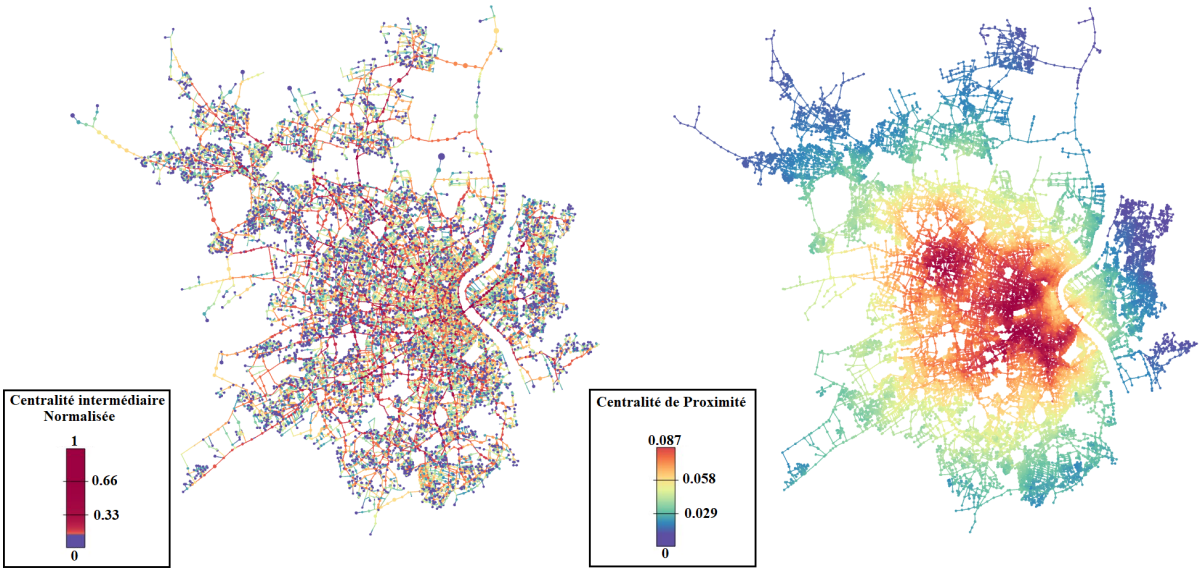


FIGURE 2.10 – Centralité intermédiaire normalisée et centralité de proximité pour le réseau *Bx*

mesure de centralité intermédiaire appliquer cette mesure dans le cas d'un graphe dirigé et pondéré. Ces contraintes de sens et de poids permettraient d'obtenir un parallèle avec la réalité hydraulique du réseau et affinerait les mesures obtenues pour la détection de points critiques et à surveiller d'un réseau AEP. Les sens et débit étant variables les mesures pourraient être obtenues sur des fonctionnements typiques ou atypiques du réseau, tel que le jour de pointe ou des débits de nuit.

2.4.2 Propriétés spectrales du graphe

Une approche courante du partitionnement des graphes consiste à exploiter les résultats de la théorie spectrale des graphes qui associent la connectivité d'un graphe à l'analyse propre de certaines matrices telles que la matrice d'adjacence $\boldsymbol{A}$ ou la matrice Laplacienne $\boldsymbol{L}$ [Chung, 1997].

Nous nous intéressons ici à quelques propriétés de la matrice Laplacienne $\boldsymbol{L}$ présentées section 2.2.4, en particulier de ses valeurs propres et vecteurs propres, en rapport avec la structure d'un graphe G.

La Figure 2.11 montre les valeurs propres $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{N_v}$ ordonnées de la matrice laplacienne L du graphe d'*Ambès*.

Un résultat formel en théorie spectrale des graphes stipule qu'un graphe G est composé de K composantes connexes si et seulement si $\lambda_1 = \cdots = \lambda_K = 0$ et $\lambda_{K+1} > 0$. Voir [Godsil and Royle, 2001] Lemme 13.1.1.

Dans notre cas, les graphes d'*ambès* et *Bx* sont deux graphes connexes pour lesquels $\lambda_1 = 0$. La seconde plus petite valeur propre λ_2 est appelée valeur de Fiedler, et le vecteur propre associé ϕ_2 le vecteur de Fiedler (voir [Fiedler, 1973]). Plus la seconde valeur propre λ_2 est grande plus le graphe G est connecté et plus il est difficile de le séparer en sous-graphes connexes en éliminant un faible nombre d'arêtes.

Le vecteur de Fiedler est couramment utilisé dans les méthodes de partitionnement

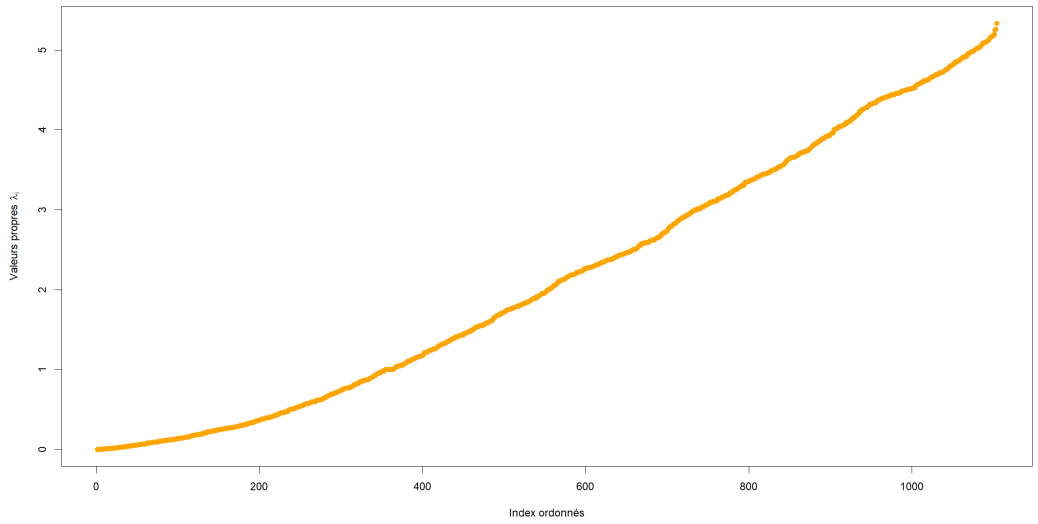


FIGURE 2.11 – Valeurs propres ordonnées λ_i de la matrice Laplacienne L du réseau *Ambès*

spectral de graphe. L'idée du partitionnement spectral est de trouver une *valeur de division* s et une partition des nœuds de G de la forme :

$$S = \{v \in V : \phi_2(v) \geq s\} \text{ et } \bar{S} = \{v \in V : \phi_2(v) < s\} \tag{2.4.1}$$

Il existe plusieurs choix populaires pour la valeur de division s en utilisant la médiane, le plus grand saut dans les valeurs des entrées du vecteur propre, ou encore le meilleur ratio de la coupure (2.4.2). Fiedler suggère un partitionnement des nœuds en fonction du signe de leurs entrées dans le vecteur propre correspondant ϕ_2. Ces méthodes de partitionnement de graphe fonctionnent particulièrement bien lorsque les degrés du graphe G sont bornées [Spielman and Teng, 2007]. La Figure 2.12 fournit une représentation visuelle du partitionnement spectral obtenue à partir du vecteur de Fiedler et de la *valeur de division* $s = 0$ pour la graphe *Bx*.

Avec des nœuds de degré maximal $\Delta(v) = 4$ pour *Bx* et $\Delta(v) = 3$ pour *Ambès*, les graphes sont creux et non fortement connexes. Il en résulte qu'une grande partie des valeurs propres sont faibles, avec une grande partie des premières valeurs propres inférieures à 1 pour le graphe d'*Ambès*.

Il apparaît alors que l'on peut facilement partitionner le graphe en K composantes en supprimant un faible nombre d'arêtes. La seconde plus petite valeur propre λ_2 suggère qu'il existe une bissection du graphe en $K = 2$ composantes ; la troisième λ_3 en $K = 3$ composantes, etc.

La Figure 2.13 fournit une représentation des vecteurs propres ϕ_i associés aux neuf plus petites valeurs propres non nulles. Les valeurs négatives de chaque vecteur propre sont représentées en bleu et les positives en rouge. L'aire de chaque nœuds est proportionnelle à l'ampleur de sa valeur dans le vecteur propre correspondant.

Pour chacun de ces vecteurs propres, on observe qu'il y a des groupes de nœuds avec des signes et ampleurs similaires, indiquant un comportement lisse sur le graphe. Et à mesure que les valeurs propres augmentent les vecteurs propres deviennent de moins en moins lisses.

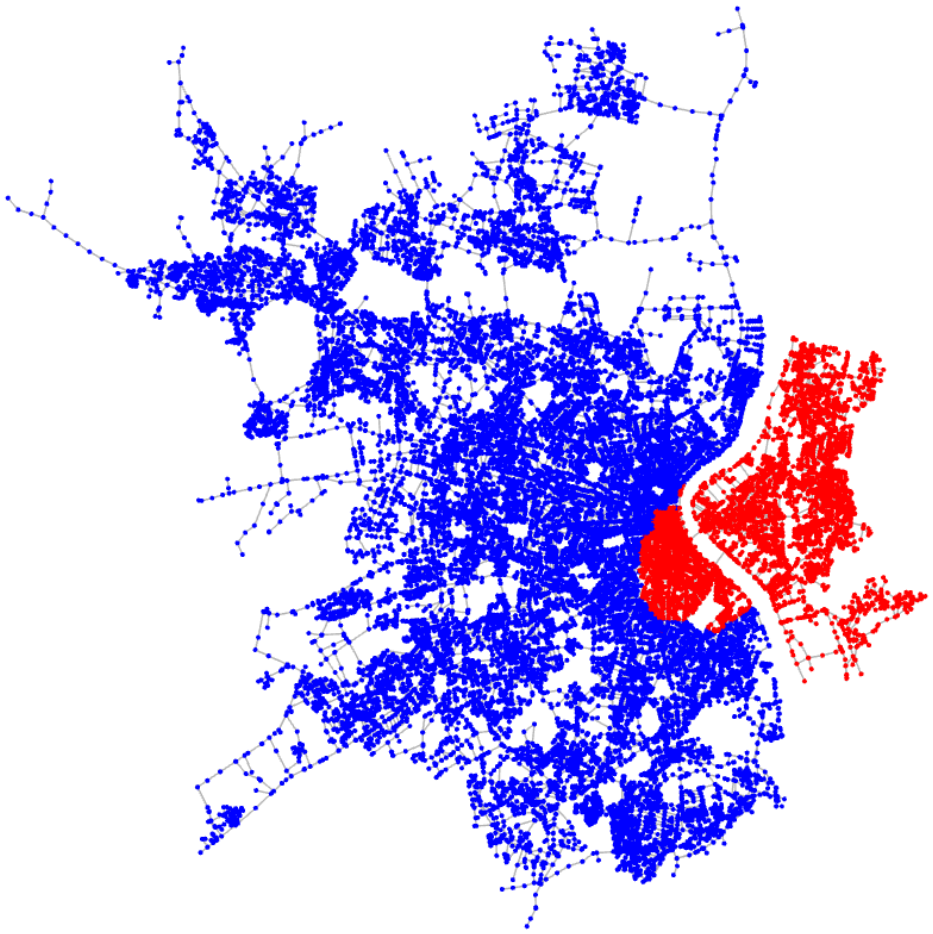


FIGURE 2.12 – Partitionnement spectrale du graphe de Bx. En rouge les nœuds appartenant à $S = \{v \in V : \phi_2(v) \geq 0\}$ et en bleu à $\bar{S} = \{v \in V : \phi_2(v) < 0\}$

On observe assez facilement sur la représentation du vecteur propre associé à la seconde plus petite valeur propre λ_2 la bissection du graphe obtenue avec la *valeur de division* $s = 0$. Les nœuds en bleu et en rouge suggérant chacun l'appartenance à une composante. Pour chaque vecteur propre on peut observer une proposition de découpe du graphe en $K = 2, 3, \ldots, 10$ composantes obtenue en supprimant un faible nombre d'arêtes. Il est intéressant de constater que la découpe en $K = 5$ composantes est très proche de la division du réseau AEP d'Ambès en cinq secteurs hydrauliques (Figure 2.15) à ceci près que la partie inférieure droite du réseau fait partie d'un seul et même secteur hydraulique et que l'analyse spectrale suggère une découpe de cette zone en deux composantes.

L'analyse spectrale d'un graphe nécessite l'obtention des vecteurs propres et valeurs propres de celui-ci. Il est donc nécessaire d'effectuer la décomposition en element propre de la matrice Laplacienne L de taille $N_v \times N_v$. Cette décomposition est couteuse en temps $\mathcal{O}({N_v}^3)$ et en ressources demandées lorsqu'il s'agit de grand graphe comme celui de *Bx*.

Néanmoins si l'on souhaite seulement se concentrer sur le partitionnement spectral d'un graphe seul le vecteur de Fiedler ou les K plus petits vecteurs propres associés aux plus petites valeurs propres sont nécessaires. Les travaux de [Mihail, 1989] par exemple montrent que pour obtenir la meilleure coupe d'un graphe planaire régulier, une approxi-

mation du vecteur de Fiedler suffit.

Plusieurs méthodes computationnelles existent afin d'obtenir de tels vecteurs sans effectuer la décomposition en éléments propres complète d'une matrice de grande de taille. Afin d'obtenir les K plus petits vecteur propres du réseau de graphe Bx composé de plus de $N_v = 60000$ nœuds nous avons utilisé le package *RSpectra* sur R, basé sur des algorithmes de Lehoucq et Sorenson ([Lehoucq et al., 1998])*. La figure en Annexe F fournit une représentation visuelle des vecteurs propres ϕ_i associés aux 9 plus petites valeurs propres λ_i de $\boldsymbol{L}$ sur le grand graphe de Bx.

*. (voir aussi la thèse de Lehoucq : [Lehoucq, 1995])

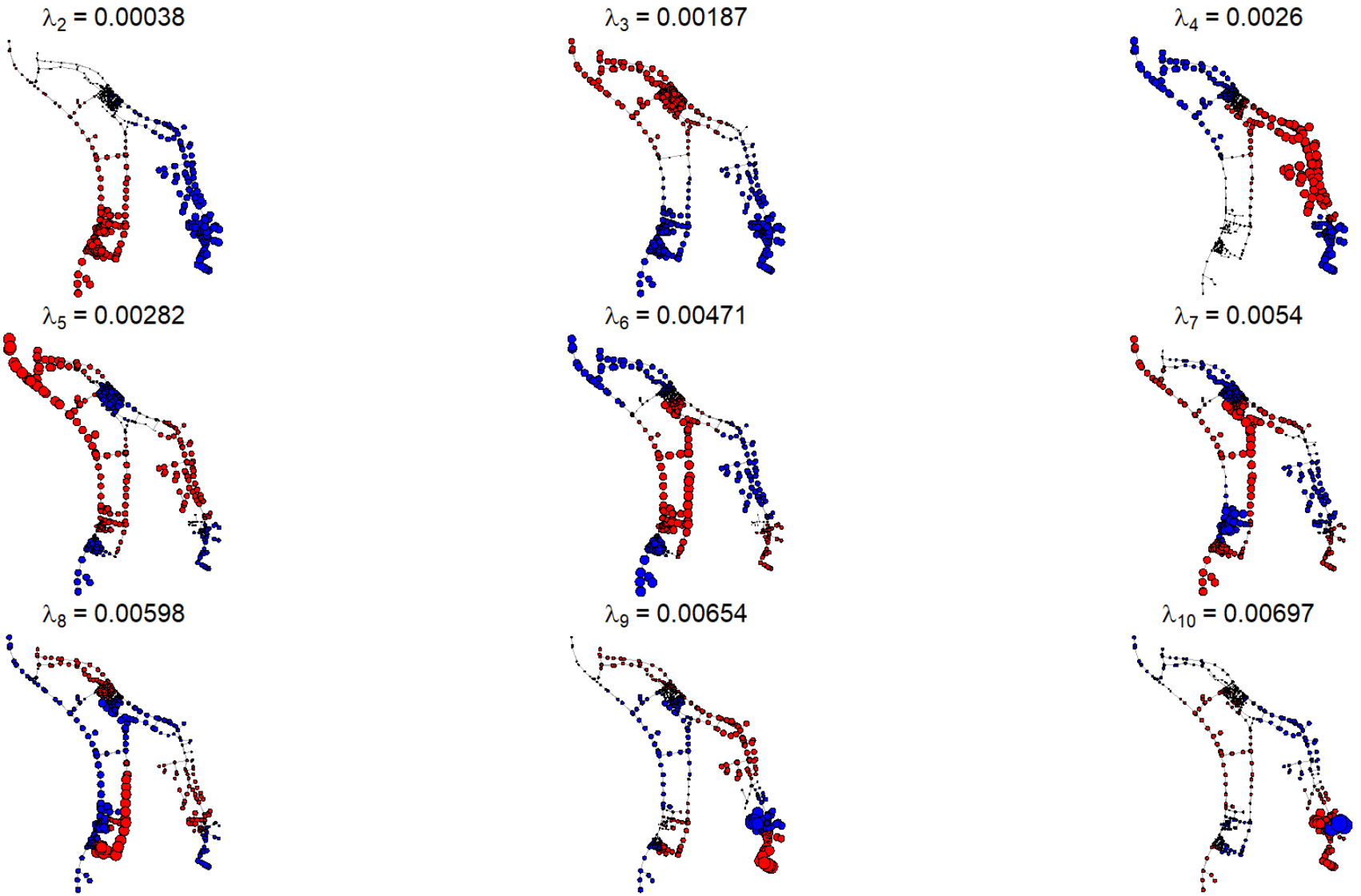


FIGURE 2.13 – Représentation visuelle des vecteurs propres ϕ_i associés aux 9 plus petites valeurs propres λ_i de L du réseau *Ambès*. Rangée supérieure : $i = 2, 3, 4$; rangée du milieu : $i = 5, 6, 7$; rangée du bas : $i = 8, 9, 10$. Les valeurs négatives sont représentées en bleu et les positives en rouge, avec l'aire de chaque nœud proportionnelle à l'ampleur de sa valeur dans le vecteur propre correspondant.

2.4.3 Proposition de nouveaux outils pour la conception et la gestion patrimoniale

2.4.3.a Agrégation de scores pour la création de chantier de renouvellement

Les réseaux d'assainissement n'ont pas fait l'objet d'étude approfondie dans cette thèse, nous avons tout de même pu transposer la structure réseaux complexe aux réseaux d'assainissement. Nous présentons ici la création d'un module pour la constitution de chantier de renouvellement d'un réseau d'assainissement. Le module présenté ici est appliqué sur un réseau de graphe représentant le réseau d'assainissement de la Métropole Bordelaise mais pourrait être appliqué à n'importe quel type de réseau.

La gestion patrimoniale d'une infrastructure consiste à la maintenir en état, tout au long de son cycle de vie, pour optimiser le coût des opérations d'acquisition, d'exploitation ou de réhabilitation afin de fournir un niveau de service performant.

Les réseaux d'assainissement, enterrés et non visitables, connaissent une dégradation structurelle, hydraulique ou d'étanchéité qu'il est nécessaire de surveiller et d'inspecter.

Malheureusement l'inspection de la totalité du linéaire n'est pas envisageable à court ou moyen terme. Des outils d'analyse ont donc été mis en place afin de simuler la dégradation du patrimoine réseau à partir des données disponibles et des observations d'une partie du patrimoine. Les résultats de ces outils permettent d'organiser des plans opérationnels de renouvellement afin de conserver le réseau le plus à jour et fonctionnel possible.

L'outil PREVOIR® Assainissement est un outil de simulation fournissant des plans théoriques de renouvellement de canalisation pour des réseaux d'assainissement. Un score de priorité est calculé pour chaque canalisation individuellement et permet de déterminer quelles sont les canalisations les plus à risque et nécessitant une inspection ou un renouvellement. Le score de priorité étant un score individuel il ne prend pas en compte l'état des canalisations adjacentes. Ainsi il n'est pas toujours évident d'identifier les chantiers les plus prioritaires.

Nous avons donc développé un module complémentaire à l'outil PREVOIR® Assainissement en vue de créer un outil expert permettant l'identification par chantier des besoins de renouvellement prioritaire sur le réseau d'assainissement.

L'objectif est de mettre en place une méthode de construction automatisée de chantiers de renouvellement fondée sur les résultats issus de PREVOIR® ainsi que de caractéristiques structurelles et opérationnelles.

Le réseau d'assainissement de la métropole bordelaise est alors représenté sous la forme d'un réseau de graphe. Le réseau $G_A(E, V)$ est composé de $N_V = 217229$ nœuds et $N_E = 215095$ arêtes. Les arêtes représentent les canalisations du réseau d'assainissement et les nœuds tous les accessoires pouvant connecter deux canalisations entre elles, par exemple des regards. Chaque arête $E \in G_A(E, V)$ dispose donc de caractéristiques structurelles ou opérationnelles permettant de décrire les règles métiers utilisées lors de la constitution de chantiers de renouvellement, par exemple :

adresse du chantier : les travaux de renouvellement nécessitant de bloquer la voirie, ils doivent être situés dans la même rue afin de minimiser l'impact sur le réseau routier ;

longueur du chantier : la longueur maximale l_{max} d'un chantier de renouvellement est fixée ;

coût maximal du chantier : chaque chantier ne peut dépasser un certain montant fixé C_{max} par la direction de l'Eau en fonction de différents critères économiques.

L'algorithme développé fonctionne en deux étapes, la première vise à constituer la liste des chantiers de renouvellement prioritaire par rue sur l'ensemble du réseau d'assainissement et la seconde étape à les hiérarchiser à l'échelle globale du réseau d'assainissement. Le niveau de priorité des chantiers est alors calculé en fonction du score de priorité issu de l'outil PREVOIR® de chaque canalisation qui les constitue.

Constituer une liste des chantiers prioritaires par rue. Un algorithme de parcours de graphe a été développé pour définir toutes les combinaisons de chaines simples $\mu(u, v)$ représentant les chantiers de renouvellement.

Afin de s'assurer que les chantiers obtenus ne s'étalent pas sur plusieurs rues, on définit un sous-graphe connexe $G_i = (V_i, E_i)$ avec $V_i \subseteq V$ et $E_i \subseteq E$ tel que toutes les arêtes E_i appartiennent à la même rue.

Un parcours en largeur (ou BFS, pour Breath-First Search en anglais) est utilisé pour déterminer toutes les chaînes simples de chacun des sous-graphes. Soit $G_i = (V_i, E_i)$ un sous graphe de G_A donné on peut alors obtenir n chaînes simples $\mu_n(u, v), \forall u, v \in V_i, u \neq v$. Étant dans un graphe pondéré on s'intéresse ici aux poids des chaînes constituées pour déterminer si celle-ci sont valides. Une chaine est considérée comme valide si :

- la longueur cumulée ne dépasse pas le critère de longueur de chantier l_{max} ;
- le coût cumulé ne dépasse pas le plafond de coût de chantier C_{max}.

1 880 037 chaînes respectant les contraintes de longueur cumulée et de coût maximum ont put être détectées avec cette méthode sur l'ensemble du graphe G_A.

Les chaînes simples obtenues représentent toutes les combinaisons de canalisation adjacente respectant les contraintes géographiques, de coût et de longueur cumulés pour chaque chantier.

Un critère intra-rue est calculé pour chacune des chaînes simples obtenues :

$$C_{intra} = \sum_{i=1}^{n} SP_i \times l_i \tag{2.4.2}$$

avec SP_i la note issue de PREVOIR® qui traduit le niveau d'urgence de renouvellement de la canalisation i, l_i la longueur de la canalisation i et n le nombre de canalisations dans la chaîne simple.

Pour chacune des rues, la chaîne disposant du critère intra-rue le plus élevé est sélectionnée et toutes les autres chaines disposant d'arêtes en commun sont supprimées de la liste. On répète l'opération pour les chaînes restantes jusqu'à ce que toutes les arêtes appartiennent à une chaîne. Ainsi toutes les canalisations d'une rue font partie d'un chantier qui maximise le critère C_{intra}. 42 056 chaines simples sont obtenues à l'aide de l'algorithme sur l'ensemble du réseau G_A.

Hiérarchiser les chantiers prioritaires Afin de hiérarchiser la totalité des chantiers nous avons défini un critère inter-rue. Ce critère a pour objectif de mettre en avant les chantiers les plus à risque en normalisant le critère intra-rue par la longueur cumulée du chantier constitué. Cela permet d'éviter l'effet d'échelle et d'accorder la priorité seulement

aux chantiers les plus longs. Le critère inter-rue est défini comme suit pour chacune des chaînes retenues précédemment :

$$C_{inter} = \frac{C_{intra}(j)}{L_j} \tag{2.4.3}$$

avec L_j la longueur cumulée du chantier j.

C'est à partir de ce score que l'on hiérarchise la totalité des chantiers obtenus.

L'ensemble des calculs ont été réalisés sur R. Les recherches de chaînes simples ont été parallélisées pour chaque rue afin de réduire les temps de calcul. L'algorithme fournit une liste hiérarchisée des chantiers prioritaires. Cette liste présente plusieurs intérêts pour l'opérateur :

— à réception de rapport d'urgence et prioritaire, pouvoir optimiser le chantier ;

— aide à la décision concernant les programmes d'inspection et de renouvellement.

Ce module complémentaire à l'outil PREVOIR$_{®}$ a permis d'améliorer la prise de décision à partir de l'utilisation des données et résultats déjà existants dans une structure de graphe adaptée aux données inter connectées et interdépendantes.

2.4.3.b Ajustement des consommations estimées par composantes connexes pour le suivi des flux par secteur hydraulique

Le suivi des bilans hydrauliques sur les réseaux de distribution peut s'avérer difficile du fait que les volumes consommés par les usagers ne sont pas mesurés en temps réel et qu'il existe en permanence sur les réseaux des volumes de fuites variables et inconnus.

Pour les compteurs n'étant pas équipés de la TLRV, on ne dispose que d'un volume totale consommé, obtenu à partir de la relève manuelle. A partir de cet historique et d'un panel de consommateurs disposant d'un niveau de consommation annuel similaire, une consommation moyenne horaire est attribuée à chacun de ces compteurs.

Cette méthode a pour effet de ne pas prendre en compte la réalité hydraulique du réseau et peut parfois sur ou sous-estimer la consommation en lissant celle-ci avec un effet moyen.

Étant donné que nous souhaitons utiliser les données de consommation pour déterminer les transferts de flux sur les réseaux depuis les sources jusqu'à chacun des points de consommation d'eau il est nécessaire de déterminer une méthode d'ajustement des consommations temps par temps. Cette méthode permettra d'ajuster spatialement les consommations en fonction des flux observés et mesurés sur le réseau à chaque instant. Pour cela nous allons utiliser le principe de bilan hydraulique permettant de déterminer les volumes entrants et sortants de secteurs hydrauliques et par déduction les volumes "consommés". Les volumes "consommés" peuvent être divisés en deux types : ceux effectivement livrés à des points de consommation et ceux que l'on considère comme perdus relevant de fuites sur le réseau. Ainsi le total des volumes facturés est inférieur aux volumes effectivement livrés au réseau.

Afin de détecter automatiquement ces secteurs hydrauliques nous allons modifier le graphe afin d'y appliquer une méthode de partitionnement utilisant le principe de composante connexe.

En théorie des graphes, le partitionnement de graphe consiste à diviser un graphe en plusieurs parties. Plusieurs propriétés peuvent être recherchées pour ce découpage, dans

notre cas nous allons vouloir découper le graphe en sous-graphes connexes pour lesquels les valeurs de flot entrant et sortant sont connues et mesurées.

Pour ce faire nous allons dupliquer et déconnecter entre eux chacun des nœuds du graphe disposant de valeurs de débit de tel que $\forall u \in V_{obs}, d(u) = 1$. De ce fait si $d^+(u) = 1, d^-(u) = 0$ alors le capteur fournit une information de flux entrant dans la CC et inversement si $d^+(u) = 0, d^-(u) = 1$ il fournit une information de flux sortant de la CC.

Si u et u' appartiennent à la même CC alors ce capteur ne fournit pas d'information de transfert de flux dans la CC.

Nous avons décidé d'employer l'algorithme de Tarjan [Tarjan, 1972] qui permet de déterminer les CC en un seul parcours en profondeur. L'algorithme prends en entrée un graphe et renvoie une partition des nœuds du graphe correspondant à ses CC.

Le principe est le suivant : on lance un parcours en profondeur depuis un sommet arbitraire. Les sommets explorés sont placés dans une pile P. Une fois tous les sommets marqués on les retire de la pile P. Les sommets retirés forment une composante du graphe. S'il reste des sommets non atteints on recommence à partir de l'un d'entre eux.

Le réseau de *Bx* est composé de $|P_1| = 51$ composantes connexes et celui d'*Ambès* de $|P_2| = 5$.

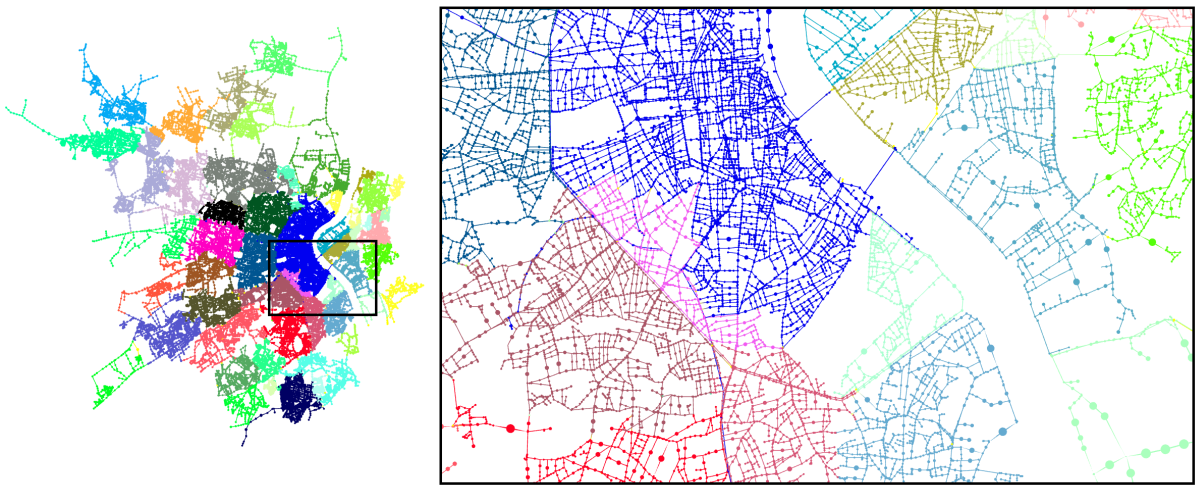


FIGURE 2.14 – Représentation sous la forme nœuds et arêtes du réseau AEP *Bx*, la couleur des nœuds et des arêtes représente l'appartenance à une composante connexe (visualisation géo-spatiale des nœuds issue du logiciel Tulip)

On peut maintenant à chaque instant déterminer pour chacune des CC les flux entrants et sortants de ses composantes comme suit :

$$\begin{cases} F^+_{p_i} &= \sum x^+(u), \forall u \in \{V_{obs} \subseteq p_i, d^+(u) = 0 \text{ et } d^-(u) = 1\} \\ F^-_{p_i} &= \sum x^-(u), \forall u \in \{V_{obs} \subseteq p_i, d^+(u) = 1 \text{ et } d^-(u) = 0\} \end{cases} \tag{2.4.4}$$

On peut donc déterminer pour chacune des CC p_i un delta représentant l'erreur liée à l'estimation de la consommation d'eau.

$$\Delta_{p_i} = F^+_{p_i} - (F^-_{p_i} + Ctot_{p_i}) \tag{2.4.5}$$

où $Ctot_{p_i} = \sum_{u \in \{V_n \subseteq p_i\}} (y(u) + \hat{y}_u)$ représente le volume total consommé dans la CC p_i, avec V_n l'ensemble des nœuds virtuels regroupant les compteurs connectés à la même canalisation et $y(u)$ le volume total mesuré des compteurs TLRV et $\hat{y}(u)$ le volume estimé des compteurs non TLRV du nœud virtuel u.

Si $\Delta_{p_i} > 0$, le volume sortant de la CC p_i est inférieur au volume entrant. Cet effet peut provenir à la fois de fuites et de sous-estimations des consommations dans le secteur. Afin de réajuster les volumes consommés on peut donc attribuer à chacun des compteurs non TLRV $\epsilon_{p_i} = \frac{\Delta_{p_i}}{|\hat{Y}_{p_i}|}$ où $|\hat{Y}_{p_i}|$ représente le nombre de compteurs non TLRV dans la CC p_i.

Si $\Delta_{p_i} < 0$, le volume sortant de la CC p_i est supérieur au volume entrant. Les consommations ont été sur-estimées et gomment probablement l'effet fuite dans le secteur. On répartit donc cette erreur de comptage sur l'ensemble des compteurs non télé-relevés pour équilibrer la balance eau fournie - eau consommée.

Dans les deux cas on fait en sorte que le volume consommé soit représentatif du volume réellement livré au réseau tel que $\Delta_{p_i} = 0$. Néanmoins il n'est pas possible avec cette méthode de détecter exactement l'origine du problème. Une des piste serait d'employer cette méthode sur une longue chronique temporelle afin de suivre l'évolution des Δ_{p_i} au cours du temps lorsque le réseau entre dans des régimes différents. Ainsi avec une analyse de ces volumes il serait possible de déterminer le caractère aléatoire ou cyclique de la variation des Δ_{p_i}.

Des fuites potentielles ou des points de livraison actifs non référencés dans les bases de données clientèles pourraient être détectés en utilisant les débits de nuit, où les consommations sont les plus basses.

La Figure 2.15 représente les 5 composantes connexes obtenues pour le graphe d'Ambès. Pour chacune des composantes connexes les valeurs de flot entrant et sortant ainsi que les valeurs de consommations estimées sont connues. Le Tableau 2.4 fournit le détail des calculs présentés précédemment pour chacune des ces composantes connexes pour le 09/12/2017 à 13h45.

La composante p_3 en bleu Figure 2.15, est desservie par une station de production d'eau potable qui envoie un débit de $20.4 m^3/_h$ dans le réseau le 09/12/2017 à 13h45. La deuxième source située dans la composante p_1 en jaune, envoie quant à elle un débit de $50.6 m^3/_h$, sur ces $50.6 m^3/_h$, $28 m^3/_h$ sont mesurés en direction de la composante p_3. N'ayant pas d'autre flot entrant dans cette composante on en déduit que $F^+_{p_3} = 48.4 m^3/_h$.

La composante p_3 est connectée à 3 composantes p_1, p_2, p_5, (respectivement en jaune, orange et rouge), par des flot sortants mesurés par trois débitmètres. $9.2 m^3/_h$ transitent vers la composante p_2, $30 m^3/_h$ vers p_5 et $1.2 m^3/_h$ vers p_1. Ainsi on peut calcul le flux sortant de p_3 tel que $F^-_{p_3} = 9.2 + 30 + 1.2 = 40.4 m^3/_h$.

Maintenant que l'on connaît les flots entrants et sortants de la composante connexe p_3 on peut calculer le volume total consommé dans cette composante connexe $Ctot_{p_3}$. La composante connexe p_3 est composée de $|Y_{p_3}| = 165$ compteurs équipés de la TLRV et de $|\hat{Y}_{p_3}| = 777$ ne disposant pas de la TLRV. Le 09/12/2017 à 13h45 il est estimé que le débit total consommé dans la composante connexe p_3 est de $Ctot_{p_3} = 31.09 m^3/_h$ avec $Y_{p_3} = 6.34 m^3/_h$ et $\hat{Y}_{p_3} = 24.75 m^3/_h$. Lorsque l'on compare cette valeur à celle obtenue à l'aide des mesures faites sur le réseau on remarque que l'on sur-estime les consommations sur cette période de la journée. En effet $\Delta_{p_3} = F^+_{p_3} - (F^-_{p_3} + Ctot_{p_3}) = -23.08 m^3/_h$.

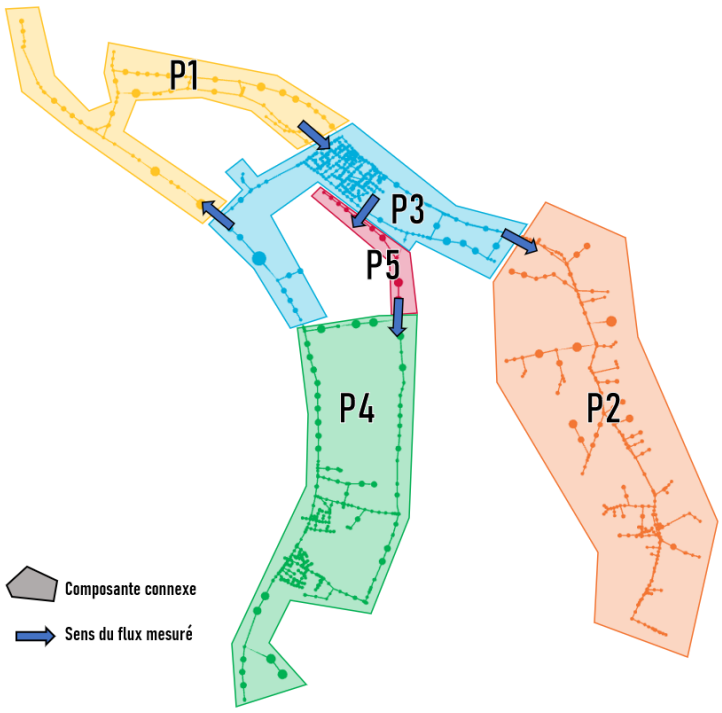


FIGURE 2.15 – Composantes connexes et sens des flux entre les composantes connexes pour le graphe d'Ambès le 09/12/2017 à 13h45

On peut donc imputer $\epsilon_{p_3} = \frac{24.75-23.08}{777} = 0.00214 m^3/_h$ à l'ensemble des 777 compteurs n'étant pas équipés de la TLRV.

La disposition des flux présentée Figure 2.15 n'est pas figée et peut être amenée à varier dans le temps. Comme expliqué Section 1.4.2.b les usines de production ont des cycles de fonctionnement, à certains moments de la journée il arrive que l'usine de production de la composante connexe p_1 ne produise pas d'eau potable, modifiant ainsi les valeurs et les sens des débits pris en compte lors du calcul des flots $F^+_{p_i}$ et $F^-_{p_i}$.

L'avantage de la méthode utilisée est qu'elle détecte automatiquement les composantes connexes du graphe et détermine de la sorte les débits à considérer comme entrants ou sortants pour chacune de celles-ci quelque soit le graphe et le pas de temps considéré.

p_i	$F^+_{p_i}$	$F^-_{p_i}$	Y_{p_i}	$\lvert Y_{p_i}\rvert$	$\hat{Y}_{p_i}$	$\lvert\hat{Y}_{p_i}\rvert$	Δ_{p_i}	ϵ_{p_i}
p_1	56.2	28	1.8567	18	3.852	115	22.4914	0.2291
p_2	9.2	0	0.1008	9	18.9775	492	-9.87	0.01849
p_3	48.4	40.4	6.3355	165	24.7452	777	-23.0807	0.00214
p_4	29.2	0	1.9145	32	29.145	744	-1.8196	0.0366
p_5	30	29.2	1.2651	4	0.2281	11	-0.6933	0.0727

TABLE 2.4 – Calcul par composante connexe des métriques permettant l'ajustement des consommations estimées pour le réseau d'Ambès le 09/12/2017 à 13h45

2.5 Conclusion du chapitre

Nous avons pu présenter dans ce chapitre la méthodologie employée pour représenter les données générées autours de la gestion d'un réseau AEP sous la forme d'un réseau de graphe. Ce réseau de graphe nous a permis de regrouper un maximum d'informations issues des bases de données et outils utilisés par le groupe SUEZ dans la gestion du réseau AEP de la métropole bordelaise.

Nous avons pu montrer qu'avec l'aide de la structure de graphe et une bonne sélections de données il était possible de répondre à des problématiques métiers concrètes et améliorer quelques processus internes déjà existants.

Il est aussi possible à l'aide de la théorie des graphes d'effectuer des analyses poussées de la topologie du réseau et ainsi mieux appréhender et identifier les potentiels points critiques ou d'intérêts de ceux-ci.

Nous avons pu durant cette thèse représenter la totalité du réseau AEP de la métropole bordelaise sous la forme d'un tel réseau. Malgré les difficultés inhérentes à des données aussi hétérogènes et parfois difficilement accessibles nous avons pu collecter un maximum de données pouvant être générées ou collectées autour de la gestion d'un réseau AEP. Toutes ces données ont pu être regroupées dans cette structure unique afin de répondre à nos problématiques. Beaucoup de freins liés à l'acquisition des données sont néanmoins encore présents.

En effet la structure constituée durant la thèse fournit une représentation topologique figée du réseau AEP. Ce biais était nécessaire afin de s'assurer que les données métrologiques utilisées soient associées à un état structurel du réseau le plus représentatif de la période de mesure. Les données utilisées ont donc été validées par des experts métier afin de s'assurer de la position et l'existence des points de mesures par exemple.

Beaucoup d'informations ne remontent pas en temps réels et nécessitent encore l'aval d'un expert pour validation. Il n'existe pas de centralisation complète de la donnée. Certaines informations sont utiles à un service en particulier et sont donc seulement présentes dans leurs bases de données.

Si l'on souhaite obtenir une reproductibilité complète du processus de création d'une telle structure de données on doit être en mesure de prendre en compte l'ensemble des données pouvant altérer ou avoir une incidence sur la structure du réseau et son fonctionnement en temps réel. Lorsqu'une vanne est fermée à 14h45 on doit pouvoir modifier la donnée sur le nœud représentant cette vanne ; si une canalisation est remplacée ou un débitmètre en court de maintenance etc..

Une telle structure pourrait être un premier pas vers l'obtention d'un jumeau numérique des réseaux AEP, permettant de représenter au cours du temps son état structurel ainsi que tout ce qui peut être mesuré, analysé ou généré autour de sa gestion.

Chapitre 3

Algorithme de flots pour la reconstruction hydraulique d'un réseau d'alimentation en eau potable

Parmi toutes les informations circulant sur un graphe, l'une d'entre elles, capitales pour le gestionnaire, est celle du débit circulant dans le réseau. Nous présentons dans ce chapitre une application des algorithmes "d'optimisation" sur les graphes pour construire une variable numérique de débit connue en tout nœud du graphe (et donc tout point du réseau physique). Outre l'intérêt évident pour le gestionnaire, cette information reconstruite nous sera utile pour tester nos méthodes de reconstruction statistique du chapitre suivant.

Nous introduisons ce chapitre avec quelques définitions et notations liés aux réseaux de flots. On s'intéresse principalement à la méthode de Ford&Fulkerson permettant de faire circuler un flot dans un réseau depuis une source vers un puits.

Ensuite nous présentons les spécificités liées aux données d'un réseau AEP qu'il est nécessaire de prendre en compte pour adapter la méthode de Ford&Fulkerson à la reconstruction hydraulique d'un réseau AEP. Premièrement les deux réseaux considérés sont composés de plusieurs sources d'eau et d'une multitude de points de consommation qu'il est nécessaire de prendre en considération pour la résolution de l'algorithme de flots. Ensuite le flot max est à chercher sous contraintes de cohérence avec les données partielles de débitmètres positionnés en certains points du réseaux. Nous présentons alors les contraintes issues des données métrologiques et structurelles nous permettant de déterminer des points de vérification de convergence de l'algorithme vers une solution en cohérence avec les flux hydrauliques mesurés.

Nous présentons enfin les résultats obtenus sur un pas de temps donné pour les graphes d'*Ambès* et *Bordeaux* nous permettant d'obtenir une fonction de débit en tout nœud du graphe. Les données étant entachées d'erreur il est nécessaire d'ajuster certaines contraintes. La détection automatique de composantes connexes, représentant les secteurs hydrauliques du réseau AEP, nous permet de vérifier la cohérence des résultats obtenus vis à vis des bilans hydrauliques mesurés.

3.1 Réseau de flot - Algorithme de flot maximum

3.1.1 Réseau de transport

On s'intéresse ici aux réseaux de transport et plus particulièrement à la recherche d'un flot maximum y circulant. Cette première partie est consacrée aux définitions des concepts utilisés dans la suite. Nous présentons par la suite la méthode de Ford&Fulkerson afin de résoudre le problème d'optimisation de flot maximum.

Tout d'abord, comme défini dans le domaine de la théorie des graphes ([Magnanti and Orlin, 1993] ; [Heineman et al., 2008]), un réseau de flot $G = (V, E, c)$ est un type de graphe orienté valué permettant par exemple, de modéliser la circulation de liquide dans des tuyaux, de courant dans un réseau électrique, de donnée via un réseau de communication, de voiture sur un réseau routier ou dans notre cas d'eau dans des canalisations.

Plus généralement, nous pouvons avec cet outil étudier et optimiser le déplacement d'une certaine quantité d'éléments dans n'importe quel réseau prédéfini.

L'analogie d'un tel réseau avec un réseau AEP est immédiate. On représente les arcs comme des conduites et leurs valuations comme la capacité maximale d'éléments pouvant transiter par ce conduit, c'est à dire sa capacité.

Le problème de flot max cherche donc à déterminer la quantité maximale d'éléments pouvant circuler dans l'ensemble du réseau, en respectant les capacités de chacun des conduits.

Définition 3.1.1 (Réseau de transport)**.** Soit $G = (V, E, c)$ un graphe orienté à valuations positives c. On dit que G est un réseau de transport s'il vérifie les conditions suivantes :

- il existe un sommet s appelé source qui n'a pas de prédécesseur,
- il existe un sommet t appelé puits qui n'a pas de successeur,
- on suppose que G est connexe, chaque sommet du graphe se trouve sur un chemin allant de la source vers le puits.

Dans ce contexte, la valuation d'un arc (u, v) est alors appelée sa capacité et est notée $c(u, v)$.

3.1.2 Flot dans un réseau de transport

Nous allons maintenant nous intéresser à la circulation d'éléments avec la notion de flot.

Définition 3.1.2. Soit $G = (V, E, c)$ un réseau de transport. Un flot f circulant dans le réseau G est une fonction à valeur réelle $f : V \times V \to \mathbb{R}$ qui, pour tous sommets u et $v \in G$, vérifie les 3 propriétés suivantes :

1. Contraintes de capacité : la quantité de flot f qui circule sur un arc doit être forcément inférieure ou égale à la capacité de cet arc. $f(u, v) \leq c(u, v)$
2. Conservation du flot (Loi de Kirchoff) : tout flot entrant dans un sommet doit en sortir. $\forall u \in \{V \mid u \neq s, t\} \sum_{i \in V, (i,u) \in E} f(i, u) = \sum_{j \in V, (u,j) \in E} f(u, j)$
3. Anti-symétrie : $f(u, v) = -f(v, u)$. Le flot du sommet u vers le sommet v doit être l'opposé du flot de v vers u.

Il découle de la loi de conservation de la Définition 3.1.2 ci-dessus qu'il n'y a pas de perte d'élément dans le réseau. La somme des flots sortants de la source est égale à la somme des flots entrants dans le puits. On appelle cette quantité la valeur de flot.

$$F = \sum_{u \in V,(s,u) \in E} f(s,u) = \sum_{u \in V,(u,t) \in E} f(u,t)$$

Définition 3.1.3 (Flot compatible)**.** Soit $G = (V, E, c)$ un réseau de transport avec c les capacités des arcs. On dit qu'un flot f sur G_f est compatible si sa valeur sur chacun des arcs est inférieure à sa capacité. $\forall (u,v) \in E, f(u,v) \leq c(u,v)$

Le nombre de flots compatibles étant nécessairement fini, il est facile de voir qu'il existe toujours un flot maximal dans un réseau de transport.

Définition 3.1.4 (Flot maximal)**.** Soit $G = (V, E, c)$ un réseau de transport. On dit qu'un flot f circulant dans G est maximal s'il est compatible et s'il possède la plus forte valeur de flot F parmi tous les flots compatibles.

Il existe en théorie des graphes différentes méthodes permettant de déterminer un tel flot. La revue bibliographique nous a conduit à utiliser l'algorithme d'Edmons-Karp qui est une implémentation spécifique de la méthode de Ford et Fulkerson afin de déterminer le flot maximal, que nous présentons dans la section suivante.

3.1.3 Méthode de Ford et Fulkerson

Avant de présenter la méthode en elle même nous allons introduire la notion de graphe résiduel (ou graphe d'écart). Il s'agit d'un graphe associé à un flot f compatible dans G, dans lequel nous cherchons les chemins augmentants permettant de déterminer un flot maximal à chaque itération.

3.1.3.a Graphe résiduel

Définition 3.1.5 (Graphe résiduel)**.** Le graphe résiduel $G_f = (V, E_f)$ d'un flot f dans G est un graphe possédant les mêmes sommets que G et dont les arcs vont dépendre de ceux de G ainsi que de la valeur du flot f y circulant. Plus précisément pour un arc $(u,v) \in E$, deux arcs sont construits dans le graphe résiduel selon la règle suivante :

— Si $f(u,v) < Cmax(u,v)$ on définit un arc (u,v) de capacité $c(u,v) - f(u,v)$.
— Si $f(u,v) > 0$ on définit un arc (v,u) de capacité $f(u,v)$.

Ainsi, pour un arc (u,v) dont la contrainte capacité n'est pas encore atteinte, on lui associe dans le graphe résiduel un arc de même sens avec la capacité restante. De plus, si un flot circule dans l'arc (u,v), on lui associe dans le graphe résiduel un arc de sens opposé avec la valeur de ce flot.

3.1.3.b Principe de la méthode

La méthode de Ford et Fulkerson consiste en un algorithme itératif (voir Algorithme 1) qui permet de déterminer un flot maximal dans un réseau de transport.

A partir d'un flot compatible f sur G, la première idée pour augmenter ce flot consiste à chercher un chemin P de s vers t sur G_f pour lequel tous les arcs ne sont pas saturés. On augmente alors le flot circulant sur ce chemin du minimum des capacités des arcs du chemin dans le graphe résiduel. De la sorte, on peut augmenter le flot sur P de la valeur :

$$\delta = \min_{(u,v)\in p} (c(u,v) - f(u,v))$$

On appelle capacité résiduelle c_f de l'arc (u,v) la valeur $c(u,v) - f(u,v)$ et on appelle chemin améliorant un chemin pour lequel $\delta > 0$.

Il suffit ensuite de mettre à jour l'ensemble des flots sur le chemin améliorant P de sorte que :

$$f(u,v) = \begin{cases} f(u,v) + \min_{(u,v)\in P} c_f(u,v) & \text{si} \quad (u,v) \in E \\ f(u,v) - \min_{(u,v)\in P} c_f(u,v) & \text{sinon} \end{cases}$$

Algorithme 1 Méthode de Ford et Fulkerson

Entrée:
Soit $G = (V, E, c)$ un réseau de transport, c les capacités
$G_f = (V, E_f, c_f)$ le graphe résiduel associé à G
Sortie: Un flot f de s à t de valeur maximale
initialiser : $\forall (u,v) \in E, f(u,v) = 0$
tant que il existe un chemin p de s à t, tel que $c_f(u,v) > 0, \forall (u,v) \in p$ **faire**
Trouver $c_f(p) = min\{c_f(u,v), (u,v) \in P\}$
pour tout $(u,v) \in P$ **faire**
1. $f(u,v) \leftarrow f(u,v) + c_f(p)$ (Envoie du flux le long du chemin)
2. $f(v,u) \leftarrow f(v,u) - c_f(p)$ (Le flux peut être "renvoyé" ultérieurement)

fin pour
fin tant que

Bien que le résultat du flot maximal tel qu'il est définit peut être intéressant. Par exemple à des fins de pilotage, pour déterminer s'il est possible de répondre à une nouvelle demande en eau dans une ville sans modifier les infrastructures. Il est nécessaire de faire quelques modifications sur le réseau de flot afin de l'adapter à la problématique de reconstruction de l'hydraulique d'un réseau.

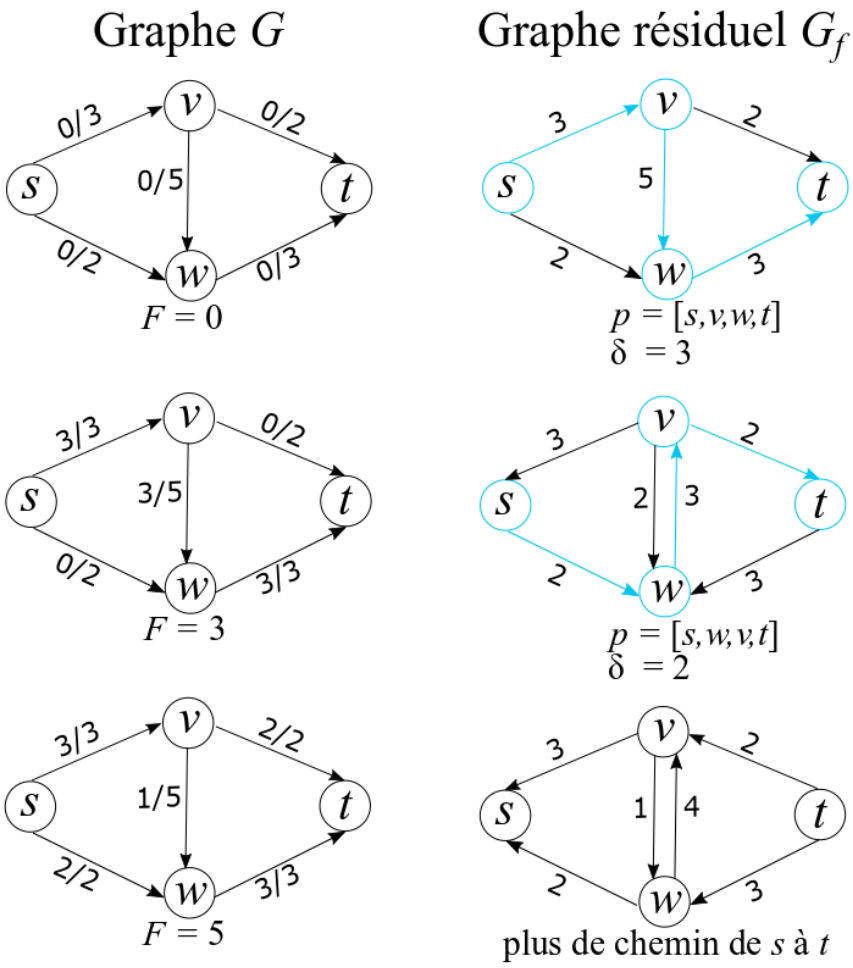


FIGURE 3.1 – Exemple de résolution de flot maximum pour un graphe G et le graphe résiduel G_f associé. F le flot actuel dans le graphe G, p le chemin sélectionné de s à t et δ le minimum des capacités des arcs sur p.

3.2 Adaptation du problème de flot maximal pour la reconstruction hydraulique d'un réseau AEP

Comme précisé dans l'introduction plusieurs ajustements ont dû être effectués pour adapter le problème du flot maximal à la fois à la problématique de reconstruction de l'hydraulique d'un réseau de flot représentant un réseau AEP mais aussi aux données disponibles.

En effet nous ne cherchons pas dans notre cas à obtenir le flot maximum compte tenu des capacités et des sens d'écoulement. Nous cherchons à déterminer les sens d'écoulement et débits sur l'ensemble des arcs compte tenu de certaines caractéristiques connues et mesurées en certains points du réseau de transport à un instant t.

Connaissant les volumes entrants et sortants, les débits et directions des flux en des

points précis, on cherche à déterminer une valeur de flot f sur $G = (V, E)$

$$f \rightarrow \sum_{i \in V^s} Cmax(s, i) = \sum_{j \in V^t} Cmax(j, t) \tag{3.2.1}$$

respectant les contraintes de direction et de capacité en des points précis du graphe. L'objectif est de déterminer les chemins empruntés par les flots ainsi que les quantités transportées sur ces chemins.

Plusieurs ajustements dans la structure du réseau de transport et dans la méthode de Ford-Fulkerson sont effectués afin de répondre à notre problématique.

3.2.0.a Réseau multi-sources et multi-puits

Contrairement au réseau de transport présenté dans la section précédente, les réseaux étudiés ici possèdent plusieurs sources et plusieurs puits. En effet nous disposons d'autant de nœuds sources $V_s \subseteq V$ que de sources d'eau dans le réseau AEP $|V_s| = 60$ et autant de nœuds puits $V_t \subseteq V$ que de points de livraison dans le réseau $|V^t| = 34000$. Comme décrit dans Cormen et al. (2009) [Cormen et al., 2009], nous pouvons réduire le problème de flot maximal avec plusieurs sources et plusieurs puits à un problème ordinaire avec un seul puits et une seule source Figure 3.2.

Pour cela, on ajoute un nœud *supersource* s et des arêtes sortantes $(s, s_i) \forall i \subseteq V_s$ pour chacun des nœuds sources. La capacité $Cmax_{(s,s_i)}$ est celle correspondant au débit d'entrée qui était mesuré sur le capteur du nœud source s_i.

Nous créons également un nouveau nœud *supersink* t et on ajoute des arêtes entrantes $(t_j, t) \forall j \subseteq V_t$ pour chacun des nœuds puits t_j. La capacité $Cmax_{(t_j,t)}$ est définie pour être le débit sortant mesuré au nœud t_j.

Intuitivement, tout flux dans le réseau multi-sources, multi-puits reste le même. La source unique s fournit autant de flux que souhaité par les sources multiples s_i et le puits unique t consomme également autant de flux que consommé par les puits multiples t_j.

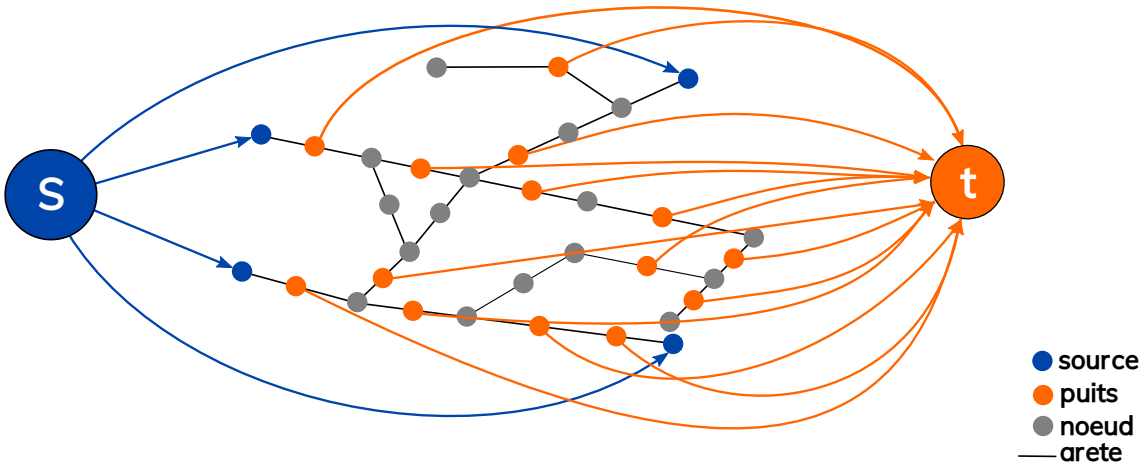


FIGURE 3.2 – Conversion d'un problème comportant plusieurs sources et plusieurs puits en un problème avec une source et un seul puits

3.2.0.b Contrainte de capacités et de sens

On souhaite définir une capacité reflétant un débit maximum, dans chacune des canalisations. Les propriétés structurelles des canalisations nous permettent de déterminer un débit théorique maximal pour chaque conduite en fonction de leurs diamètres et une vitesse de circulation maximale fixée. On définit alors la capacité $Cmax(u, v)$ de l'arête (u, v), comme étant le débit maximal théorique en m^3/h pouvant traverser l'arête (u, v).

$$Cmax(u, v) = s_{(u,v)} \times V_{max}, \tag{3.2.2}$$

avec $s_{(u,v)} = \pi \times (\frac{d_{u,v}}{2})^2$ la surface et V_{max} la vitesse maximale du flot de la canalisation (u, v).

Bien que cette capacité théorique représente, comme dans un réseau de flot, la capacité maximale ne pouvant être dépassée, elle ne nous permet pas de définir le sens d'écoulement de celui-ci.

En effet les débits et sens d'écoulement sont inconnus dans une grande partie des canalisations, le réseau de graphe le représentant est initialement non dirigé. Il est nécessaire que celui-ci soit dirigé afin de déterminer des chemins augmentant de la source s au puits t. De ce fait nous dupliquons toutes les arêtes pour créer des arcs aller et retour dans le réseau.

Pour chaque arête $(u, v) \in E$, nous créons une arête supplémentaire afin d'obtenir deux arcs antiparallèles (u, v) et (v, u) et nous définissons la capacité des deux arcs sur celle d'origine $Cmax(u, v) = Cmax(v, u)$. A la différence d'une résolution classique de problème de flot maximum avec des arcs antiparallèles, nous ne pouvons emprunter les deux arcs à la fois. En effet, le flux à l'intérieur d'une canalisation ne peut pas aller dans deux directions à la fois ; il est donc interdit de laisser le flux se trouver simultanément dans les deux arcs antiparallèles.

Cette manipulation nous permet d'autoriser le parcours d'une arête (u, v) dans un des deux sens dès l'initialisation de l'algorithme de flot. On peut ainsi mettre à jour l'information de l'arc emprunté par le flot en fonction de la valeur de flot. Si le flot est présent dans l'arc (u, v) alors l'arc (v, u) devient l'arc résiduel de G'. Ainsi tant que le flot dans l'arc résiduel (v, u) est inférieur ou égale à celui de l'arc (u, v), la mise à jour des capacités résiduelles et du flot est identique à celle du problème de flot maximum. Lorsque celui-ci devient majoritaire dans l'arc résiduel, l'arc (u, v) devient l'arc résiduel et inversement l'arc (v, u) devient l'arc parcouru par le flot.

Une fois que l'algorithme de flot max est terminé le sens des arêtes est déterminé par celles ayant un flot strictement positif $f(u, v) > 0$

$$c'(u, v) = \begin{cases} Cmax(u, v) - f(u, v) & si (u, v) \in E \\ f(v, u) & si (v, u) \in E \\ 0 & sinon \end{cases} \tag{3.2.3}$$

3.2.0.c Prise en compte des observations

Comme présenté dans la Section 1.4.2.b les débits sont connus et mesurés en certains nœuds du réseau équipés de débitmètres. On peut donc remplacer la capacité théorique

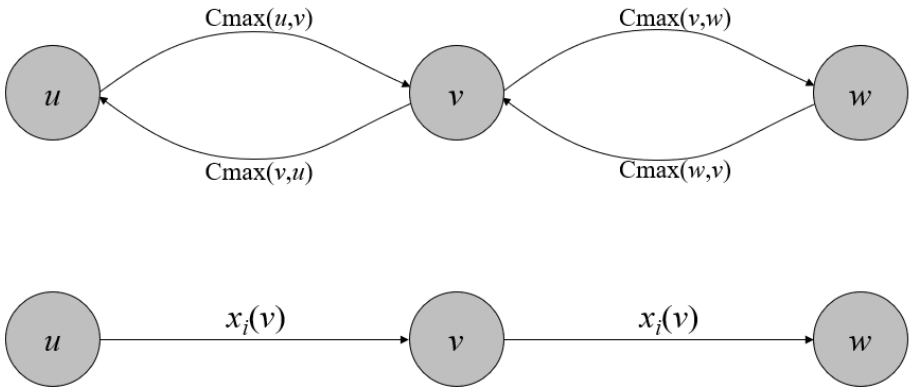


FIGURE 3.3 – Valuation et orientation des arêtes d'un réseau de flot. En haut sens d'écoulement et débits inconnus, en bas connus et mesurés

$Cmax(u, v)$ par une capacité observée et à atteindre par l'algorithme de flot pour toute arête voisine d'un nœud observé.

On différencie deux types de nœuds observés :

— Les nœuds voisins de la source ou du puits qui fournissent des informations sur les entrées ou les sorties de flux dans le réseau. s, $V_s = \{1, \ldots, n\}, n = 60$ pour les flux entrants et $V_t = \{1, \ldots, m\}, m = 34000$ pour les flux sortants ; respectivement en bleu et orange Figure 3.4.

— Les nœuds équipés d'un débitmètre qui fournissent une mesure du débit dans le réseau $V_{obs} = \{1, \ldots, i\}$. $Cmax(u, v) = x_{obs}(u), \forall u \in V^{obs}$, avec $x_{obs}(u)$ le débit mesuré par le débitmètre au nœud u en vert sur la Figure 3.4.

La direction du flux pour l'ensemble de ces nœuds est connue. De cette façon nous pouvons directement créer un arc représentant le sens du flot mesuré. Il n'y a donc pas d'arête bidirectionnelle et il est de plus interdit de faire réduire le flot en passant par ces arcs. Le flot ne peut qu'augmenter car nous cherchons à faire converger le flot dans ces arcs vers le débit mesuré. Contraindre l'augmentation de la valeur de flot sur ces arcs entraine le fait que la solution fournie par l'algorithme n'est pas le flot maximal du réseau de BM, mais le résultat maximum au regard des contraintes.

3.2.0.d Recherche de chemin augmentant

La méthode de Ford-Fulkerson laisse libre choix quant à la recherche de chemin optimal de s à t à chaque itération de l'algorithme de flot maximal. Intuitivement nous cherchons à obtenir les chemins les plus courts pouvant séparer la source du puits. De ce fait nous devons à chaque itération déterminer un chemin simple $p = \{s, v_1, ..., v_n, t\}$. Le nombre de chemins élémentaires entre s et t étant fini il est possible de tous les énumérer (Lemme de Kőning [Kőnig, 1927]). Il existe des algorithmes de recherche permettant d'éviter la combinatoire d'explorer tous les chemins élémentaires d'un graphe. Au départ tous les sommets sont non marqués, à l'exception du sommet de départ s. A chaque étape un nœud non marqué adjacent à un nœud marqué est sélectionné et marqué à son tour. Le choix d'un sommet voisin d'un sommet déjà marqué assure que l'ensemble des sommets marqués est un sous-graphe connexe à chaque étape de l'algorithme.

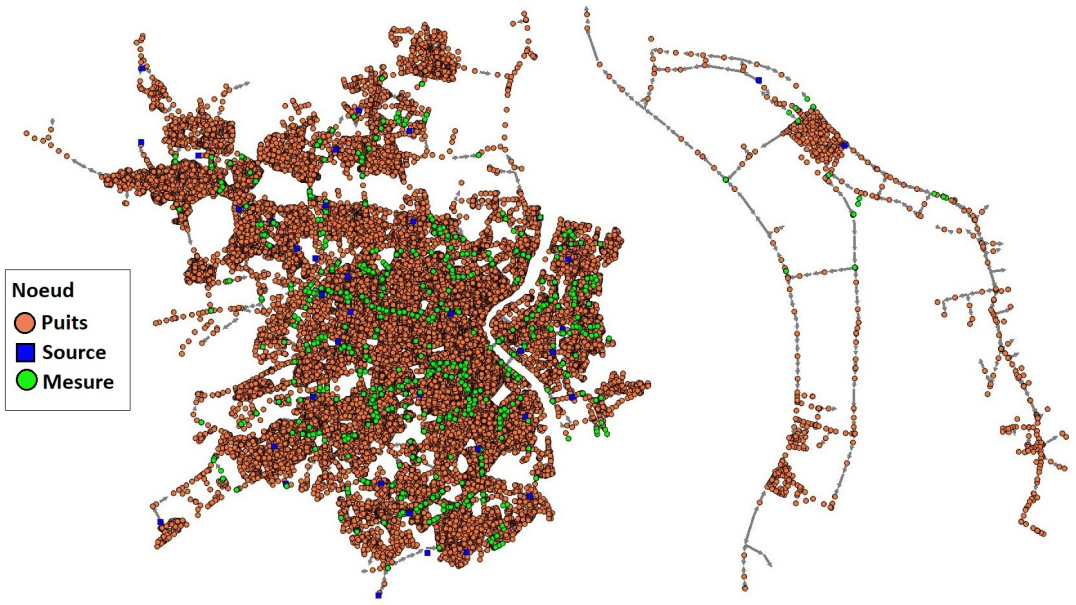


FIGURE 3.4 – Représentation des réseaux de transport *Bx* à gauche et *Ambès* à droite, pour plus de lisibilité les nœuds intermédiaires ne sont pas affichés. Seuls les nœuds disposant de mesures sont affichés.

Il existe différentes stratégies permettant de sélectionner à chaque étape le sommet à marquer : la recherche en profondeur (ou DFS, *Depth First Search* en anglais) et la recherche en largeur (ou BFS, *Breath First Search* en anglais). Comme dans la méthode Edmonds-Karp [Edmonds and Karp, 1972], nous utilisons un algorithme de recherche par largeur (BFS) pour trouver les chemins augmentant. Contrairement à la recherche en profondeur dont le but est d'aller très vite profondément dans le graphe, la recherche en largeur consiste à épuiser la liste des sommets proches du nœud de départ s avant de poursuivre plus en profondeur.

Nous souhaitons contraindre l'algorithme à trouver les chemins augmentant passant par les sommets observés pour saturer les flux dans les arêtes adjacentes. On considère alors un sous-ensemble d'arêtes $x_{obs} = (u, v)$ pour tout $u \in V^{obs} et\ v \in \Gamma(u)$ contenant toutes les arêtes incidentes à un nœud observé, pour lesquelles les débits et sens d'écoulement sont connus. Parce qu'elles représentent une observation réelle de flux dans le réseau, on choisit de les saturer en premières tout en résolvant le problème de flot maximal. Cette contrainte entraîne une solution maximale contrainte, qui diffère de la solution de flot maximal donnée par la méthode de Ford-Fulkerson, comme l'illustre la Figure 3.5.

Un BFS bidirectionnel est utilisé pour s'assurer que le chemin obtenu aille d'abord de la *supersource* s jusqu'à un des nœuds observés x_{obs} puis du nœud observé marqué jusqu'au *superpuits* t. Comme décrit dans la Figure 3.6, on commence au nœud source et explore les nœuds voisins avant de passer au niveau voisin suivant. Nous répétons le même processus jusqu'à atteindre le nœud observé $i \in V^{obs}$. On répète ensuite l'opération depuis i jusqu'au puits t. Les deux chemins obtenus doivent former un chemin élémentaire, aucun circuit n'est autorisé. Si les deux chemins ne sont pas vides, nous avons trouvé un chemin valide de s à t via le nœud de capteur $p = s \rightarrow x_{obs} \rightarrow t$.

Le principe de sous-optimalité nous garantit que le chemin obtenu de la sorte est

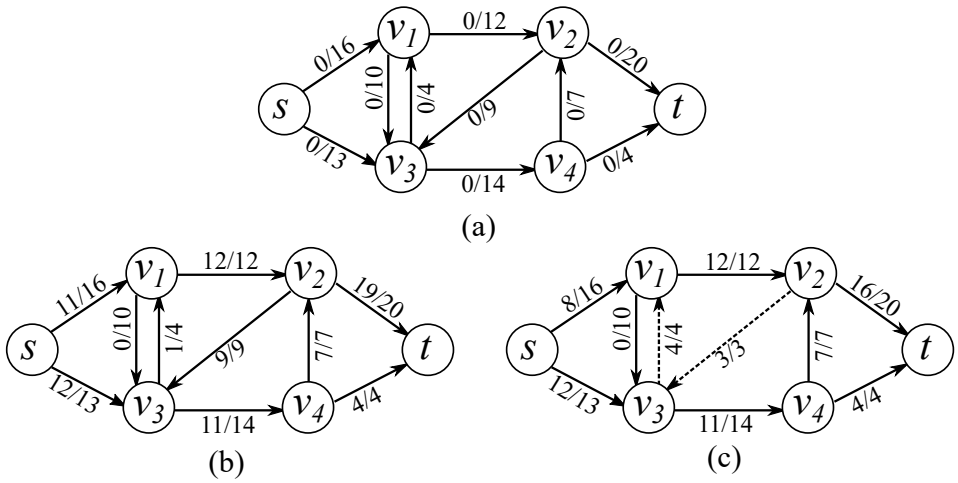


FIGURE 3.5 – (a) Un réseau de flot $G = (V, E)$ avec une source s et un puits t. Chaque arête est étiquetée avec son débit et sa capacité (flot/capacité). (b) Un flux f_1 dans G de valeur maximal $|f_1| = 23$. (c) Un flux f_2 dans G de valeur $|f_2| = 20$ respectant les contraintes de sens et de capacité des arêtes observées (v_3, v_1) et (v_2, v_3).

toujours le chemin le plus court [Bellman, 1957]. Il s'agit de remarquer que la notion de plus court chemin est héréditaire.

Propriété 3.2.1. *Si $p = \{s, ..., t\}$ est un plus court chemin entre s et t alors pour tout sommet x sur le chemin p*

Le sous-chemin de s jusqu'à x, $\{s, ..., x\}$ est un plus court chemin de s à x.

Le sous-chemin de x jusqu'à t, $\{x, ..., t\}$ est un plus court chemin de x à t.

Démonstration. Si un des sous-chemins de s jusqu'à x ou de x jusqu'à t n'était pas un plus court chemin alors on pourrait le remplacer par un chemin plus court pour obtenir un chemin de s à t plus court que p : absurde. □

Tant qu'un des deux chemins n'est pas saturé il reste considéré comme étant le plus court et valide pour l'algorithme de flot maximal. Il est donc inutile de recalculer un tel chemin. La recherche en largeur combinée permet à la fois de s'assurer que les nœuds disposant d'observations font partie des plus courts chemins élémentaires sélectionnés et de réduire les temps de calculs en ne recalculant que les chemins ayant été saturés par un flot.

Une fois l'ensemble des nœuds disposant d'observations saturé, c.à.d une fois qu'il n'est plus possible d'augmenter le flot passant par ces nœuds, nous effectuons une recherche classique de chemins augmentant de la source s jusqu'au puits t, tant qu'il existe des chemins valides.

Algorithme 2 Méthode de Ford et Fulkerson contrainte

Entrée:
Soit $G = (V, E, c)$ un réseau de transport, c les capacités
$G_f = (V, E_f, c_f)$ le graphe résiduel associé à G
Sortie: Un flot f de s à t de valeur maximale
tant que il existe un chemin **faire**
si tous les noeuds observés sont non visités ou non saturés **alors**
$p_1 \leftarrow$ chemin de s à x_{obs}
$p_2 \leftarrow$ chemin de x_{obs} à t
$p \leftarrow$ chemin de s à t en passant par x_{obs}
sinon
$p \leftarrow$ chemin de s à t
fin si
$c_f(p) = min\{c_f(u, v), \forall (u, v) \in p\}$
pour toute arête $(u, v) \in p$ **faire**
$e_0 \leftarrow$ première arête (u, v)
$e_1 \leftarrow$ seconde arête (v, u)
si $f(u, v) > 0$ **et** $f(v, u) = 0$ **alors**
$f(u, v) \leftarrow f(u, v) + c_f(p)$
sinon
$f(v, u) \leftarrow f(u, v) + max(0, c_f + f(u, v))$
$f(u, v) \leftarrow max(0, f(u, v) - c_f(p)$
fin si
fin pour
fin tant que

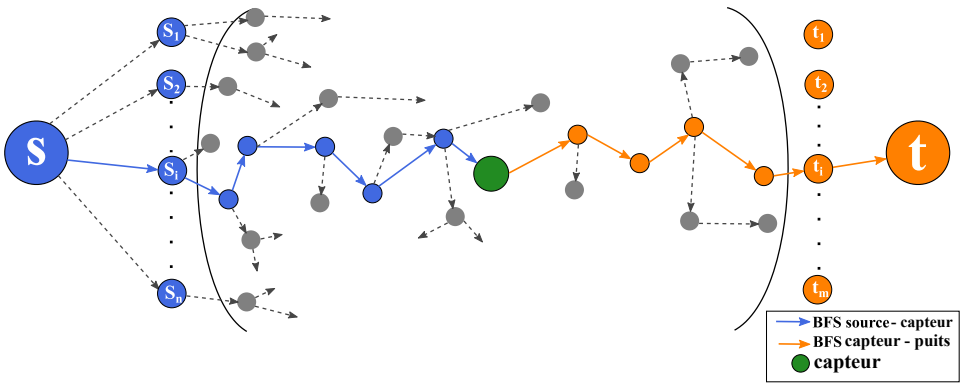


FIGURE 3.6 – Parcours bidirectionnel en largeur : du nœud source s au nœud observé v et de v au nœud puits t

3.3 Reconstruction de l'hydraulique du réseau à partir des données mesurées

Nous décrivons dans cette partie les résultats obtenus par l'utilisation du modèle de graphe orienté données (MOD), pour la reconstruction de l'hydraulique d'un réseau AEP (Section 3.2).

Nous choisissons une implémentation en *C++* de la méthode Ford-Fulkerson [Fulkerson and Ford, 1962] pour calculer le problème flot maximum. Nous avons utilisé la librairie de graphe Wulip développée par le LaBRI, afin de reconstruire l'hydraulique des réseaux *Bx* et *Ambès*. Comme décrit Section 1.4 et Section 2.3.1, les deux réseaux étant non connexes, nous avons décidé de les traiter indépendamment.

Les premiers résultats obtenus à l'aide du MOD nous permettent à la fois de détecter et corriger les erreurs présentes dans les données brutes ainsi que d'améliorer la fonction des débits en tout point du réseau obtenue. Nous présentons aussi les limites actuelles d'un tel modèle liées à la qualité et quantité de données disponibles.

Les deux réseaux de transport $G_1 = (V_1, E_1, c_1)$ du réseau *Bx* et $G_2 = (V_2, E_2, c_2)$ du réseau *Ambès*, sont représentés Figure 3.4.Le graphe G_1 est composé de $N_V = 91977$ nœuds et $N_E = 98172$ arêtes. Parmi ces nœuds $|V_s| = 60$ sont des nœuds sources et $|V_t| = 35000$ des nœuds puits et $|V_{obs}| = 200$ les points de mesures représentant les débits mesurés. Le graphe G_2 quant à lui est composé de $N_V = 1087$ nœuds et $N_E = 1188$ arêtes avec $|V_s| = 60$ nœuds sources et $|V_t| =$ nœuds puits et $|V_{obs}| = 5$.

Nous avons appliqué sur chacun de ces deux graphes l'algorithme de Ford-Fulkerson contraint présenté dans la section précédente (voir algorithme 2). La Figure 3.7 represente l'état de saturation obtenue pour l'ensemble des points de mesure sur les graphes de G_1 de *Bx* et G_2 d'*Ambès*.

Cette saturation représente la valeur flot à atteindre lors de l'application de l'algorithme pour chacun des nœuds disposant de mesures. Elle nous permet de déterminer la qualité de convergence de l'algorithme. Tant que ces contraintes ne sont pas satisfaites on pourra alors déterminer que l'algorithme ne fournit pas des débits et sens d'écoulement représentatifs de ceux mesurés par l'ensemble des capteurs :

Noeud source : $f(s,i) < C_{max}(s,i), \forall i \in V_s$ le flot entrant attendue n'est pas atteint

pour le nœud source $i \in V_s$. Il n'existe plus de chemin augmentant partant de i.

Noeud puits : $f(j,t) < C_{max}(j,t), \forall j \in V_t$ le flot sortant attendue n'est pas atteint pour le nœud puits $j \in V_t$. Il n'existe plus de chemin augmentant menant à j.

Débimètres : $f(u,v) < C_{max}(u,v), \forall u \in V_{obs}, v \in \Gamma(u)$ le flot passant par le débitmètre n'est pas atteint pour le nœud $u \in V_{obs}$. Lors de la saturation des nœuds observés aucun chemin n'a été trouvé permettant d'augmenter le flot sur les cas voisins du débitmètre.

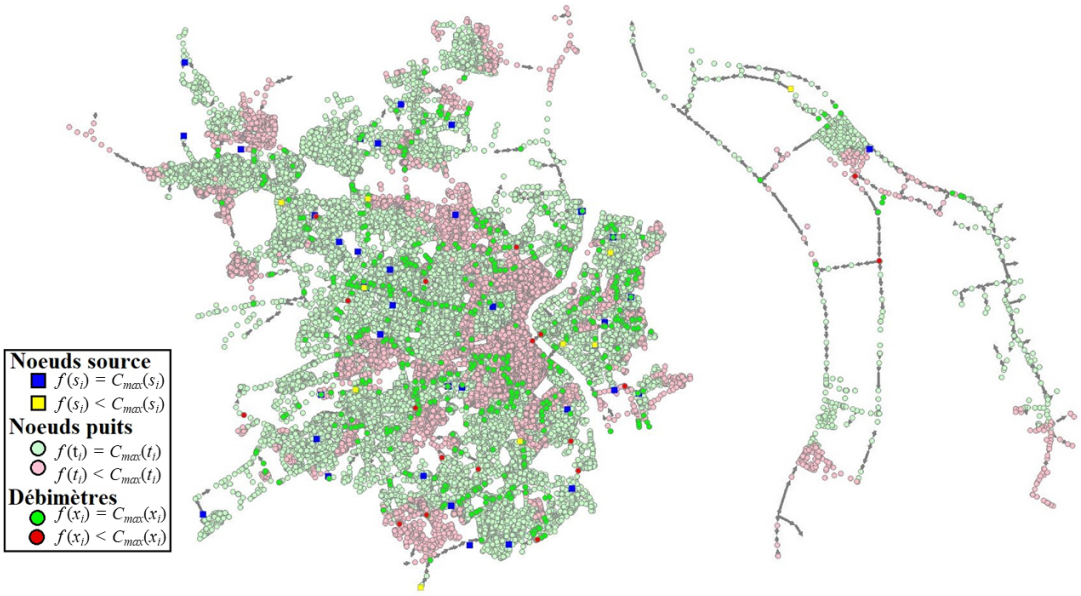


FIGURE 3.7 – Représentation graphique des résultats de convergence des contraintes, pour le graphe G_1 de *Bx* à gauche et G_2 d'*Ambes* à droite.

3.3.1 Non convergence des contraintes

La non convergence de certaines des contraintes découle du fait que la disponibilité des données et leurs qualités sont discutables à plusieurs niveaux. Le manque de connaissance des volumes consommés en temps réel ainsi que l'écart entre l'état structurel du réseau et sa représentativité dans les bases de données en sont la raison principale.

La quantité de flot représente le volume livré au réseau (VLAR). Dans le cas de l'algorithme de Ford-Fulkerson la quantité de flot entrant est par définition égale à la valeur de flot sortant. $F = \sum_{i \in V_s} Cmax(s,i) = \sum_{j \in V_t} Cmax(j,t)$.

Néanmoins, comme présenté section 1.4.2.b, les données de consommation représentant une partie des flux sortants sont entachées d'erreurs. Le volume consommé attribué à chacun des nœuds puits, $Cmax(j,t), \forall j \in V_t$ est composé de volumes mesurés pour les compteurs équipés de TLRV $y(j)$ et de volumes estimés pour ceux non équipés de la TLRV $\hat{y}(j)$ regroupé sur le nœud j.

$$Cmax(j,t) = \sum_{i=1}^{n} y_i(j) + \sum_{i=1}^{m} \hat{y}(j), \forall j \in V_t$$

où n représente le nombre de compteurs équipés de la TLRV sur le nœud j et m le nombre de compteurs ne disposant pas de la TLRV sur le nœud j.

De plus comme tout réseau AEP, ceux de *Bx* et *Ambès* disposent de fuites. Ces fuites, pour la plupart invisibles, représentent un flux sortant en des points et de volumes inconnus.

A chaque instant t il est possible de déterminer pour chacun des points de mesures un écart entre la valeur attendue (mesurée par les capteurs) et celle obtenue à partir de l'algorithme de flot. La Figure 3.7 fournit une représentation graphique de ce type de résultat de convergence pour chacun des nœuds source, puits et débitmètre réseau. On peut ainsi observer pour chacun d'eux l'écart entre la valeur réelle, mesurée par les capteurs et celle estimée suite à l'application de l'algorithme de Ford-Fulkerson contraint (voir Table 3.1).

La Figure 3.8 présente les résultats de convergence de l'algorithme de Ford-Fulkerson contraint sur le réseau d'*Ambès* le 09/12/2017 à 13h45. Pour chacun des nœuds sources et débitmètres réseau, est affiché le résultat obtenu suite à l'exécution de l'algorithme par rapport à la valeur réelle mesurée. Les nœuds puits sont aussi représentés indiquant en vert s'il la capacité maximale est atteinte et en orange sinon.

Les débitmètres en usine de production ont mesuré $20.40 m_3/h$ et $55.00 m_3/h$ respectivement pour la source 1 et la source 2.

Seuls $32.51 m_3/h$ des $55.00 m_3/h$ mesurés par la source 2 ont pu être transmis dans le réseau. En effet si l'on regarde sur le graphe de gauche Figure 3.8, l'ensemble des nœuds puits du secteur p_3 sont saturés, ainsi que le débitmètre numéro 2 ayant mesuré un flux sortant de la composante de $28.80 m_3/h$. C'est à dire qu'il n'existe plus de chemin augmentant partant de la source 2. Néanmoins l'usine de production ayant effectivement envoyé $55.00 m_3/h$ d'eau dans le réseau, il reste $22.49_3/h$ à transmettre dans réseau.

A l'inverse la source 1 a atteint sa capacité maximale, plus aucun chemin augmentant ne peut partir de la source 1. Il reste pourtant plusieurs nœuds puits n'ayant pas été saturés. Les débitmètres 3 et 4 ne sont pas non plus saturés. Cela signifie que du flux a été mesuré transitant par ces capteurs dans le réseau et pourtant l'algorithme de flot ne trouve plus de chemin augmentant permettant de saturer ces nœuds.

Avec ces deux exemples on observe le fait qu'il est très important de connaître avec précision les quantités de flot sortant du réseau pour obtenir la saturation de toutes les contraintes.

Capteur	Composante connexe		Débit (m^3/h)		
	Entrant	Sortant	Réel	Avant ajustement	Après ajustement
$Source_1$	p_3	Ø	20.40	20.40	20.40
$Source_2$	p_1	Ø	55.00	32.51	55.00
$Débit_1$	p_2	p_3	9.20	9.20	9.20
$Débit_2$	p_3	p_1	28.00	28.00	28.00
$Débit_3$	p_5	p_3	30.00	10.00	30.00
$Débit_4$	p_4	p_5	29.20	10.00	29.20
$Débit_5$	p_1	p_3	1.20	1.20	1.20

TABLE 3.1 – Valeurs réelles mesurées par les capteurs et estimées par l'algorithme de Ford-Fulkerson avant et après l'ajustement des consommations par composantes connexes, pour le réseau *Ambès* le 09/12/2017 à 13h45

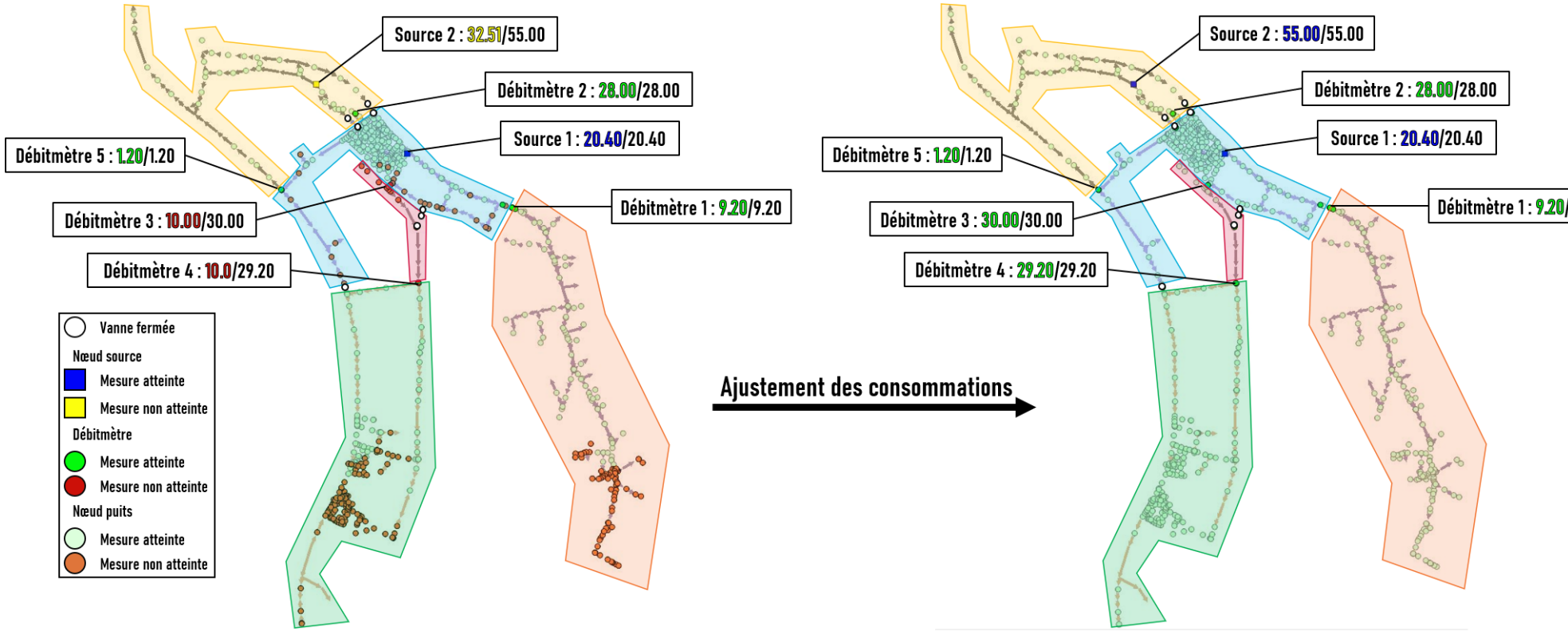


FIGURE 3.8 – Résultat de l'algorithme de Ford-Fulkerson contraint du réseau d'*Ambès* pour la journée du 09/12/2017 à 13h45, à gauche avant l'ajustement des consommations par composantes connexes et à droite après l'ajustement. Les mesures de débit des nœuds sources et des débitmètres sont indiqués sous la forme Estimé/Mesuré.

On utilise alors l'ajustement des consommations par composantes connexes, présenté section 2.4.3.b, afin d'ajuster les estimations de consommation de l'ensemble des compteurs n'étant pas équipés de la TLRV, $\hat{y}(j), \forall j \in V_t$.

Ainsi quelque soit le pas de temps étudié pour toutes les composantes connexes considérées la valeur Δ_{p_i} est calculée comme présentée Equation 2.4.5. A chacun des nœuds puits est attribuée la quantité nécessaire ϵ_i afin d'obtenir un équilibre des flux entrants et sortants.

Cette option permet alors de lisser les flots manquants sur des échelle spatiales plus fines et représentatives de l'état hydraulique du réseau à chaque instant. Néanmoins comme expliqué précédemment les réseaux ne sont pas imperméables, il est donc difficile de savoir quelle proportion de ces flux réajustés correspond à des fuites ou à des volumes réellement consommés.

Le réseau de graphe de droite Figure 3.8 présente les résultat obtenus par l'algorithme de Ford-Fulkerson contraint après avoir appliqué l'ajustement des consommations par composantes connexes. L'ensemble des nœuds disposant de mesures de débit est saturé, et la totalité des flux entrants a pu transiter jusqu'aux puits. On considère alors que la fonction de débit obtenu sur l'ensemble du graphe est représentative de l'état hydraulique du réseau ayant été mesuré par l'ensemble des débitmètres.

Le réajustement des consommations est effectué de manière homogène pour tous les compteurs n'étant pas équipés de la TLRV dans chaque composante connexe. Ceci a pour effet de lisser la répartition des flux sortants dans chaque composante. Les moments et niveaux de consommation étant hétérogènes dans la réalité ceux obtenus sont très probablement différents de la réalité hydraulique du réseau sur une échelle temporelle aussi fine ($\approx 15min$). Cet effet étant encore accentué dans les secteurs les plus maillés. Pondérer les consommations réajustées par nœuds puits en fonction de différents critères inhérents aux points de consommations permettrait de réintégrer une hétérogénéité des flux sortants.

3.4 Conclusion du chapitre et perspectives

L'algorithme présenté ici fournit une fonction des débits sur tous les nœuds du graphe. Elle est obtenue à l'aide de l'algorithme de flot maximum adapté au cas de données de flux partiellement mesurées sur les noeuds d'un graphe. Les contraintes fixées à partir des données mesurées et des données structurelles représentant le réseau de distribution nous fournissent ainsi une solution que l'on peut considérer comme "valide" au regard des données mesurées en temps réel sur le réseau. On dispose ainsi d'une fonction simulée en tout noeuds du réseau ayant un sens d'un point de vue équilibre hydraulique sur le réseau afin de tester les méthodes d'inférence statistique présentées dans le chapitre suivant.

Cet algorithme de flot pourrait être utilisé à des fins opérationnelles, comme par exemple pour le suivi dynamique de l'hydraulique du réseau. Cependant ce nouveau cas d'usage nécessiterait l'acquisition de données traduisant les évolutions structurelles et hydrauliques du réseau en temps réel.

Comme nous avons pu le voir tout long de ce chapitre, les données structurelles du réseau AEP définissent les liens entre chaque nœud et pour l'algorithme de flot les limites et chemins possibles. Néanmoins les données structurelles utilisées durant la thèse ne sont

pas dynamiques. Elles représentent une image du réseau à un instant t or la structure d'un réseau AEP n'est jamais figée. Il vieillit et subit des altérations qui peuvent modifier son fonctionnement ; qu'elles soient planifiées par des opérateurs (réparations, purges, etc.) ou imprévues (casses, fuites, etc.). Ces informations sont parfois difficilement vérifiables et ne remontent pas de façon automatique dans les bases de données. Pourtant elles permettraient de contraindre l'algorithme de flot en précisant l'état structurel du réseau à chaque instant.

Par exemple l'estimation des capacités maximales dépend du diamètre intérieur des canalisations. Cette information est renseignée dans la base de donnée au moment de la pose des canalisations. C'est une bonne indication de la quantité maximale de flot pouvant les traverser. Mais cette information ne prend pas en compte tout ce qui peut altérer l'état de la canalisation au cours du temps. Le matériau, la nature du sol dans lequel elle est enterrée, son âge, son taux d'encrassement, etc., sont autant de paramètres pouvant modifier les capacités maximales de chaque arête.

L'état d'ouverture des vannes est également une donnée qui permettrait de contraindre en certains points du réseau de flot les sens d'écoulement. Il existe sur le réseau de la métropole bordelaise plus de 2000 vannes permettant d'isoler certaines canalisations ou secteurs du réseau d'alimentation en eau potable lors d'interventions sur le réseau (réparation, purges, etc.). Se faisant certaines zones se retrouvent entièrement coupées d'alimentation en eau et aucun flux ne peut les atteindre.

Néanmoins l'actionnement des vannes n'est ni suivi ni mis à jour dans la base de données durant les interventions. Ainsi si une vanne est fermée pendant 2 heures son état temporaire n'est pas modifié dans la base de données. Le suivi en temps réel de l'état des vannes permettrait d'interdire ou non le passage du flot par certaines arêtes ou sous-graphe avec certitude.

Enfin le manque d'informations des volumes sortants sur des pas de temps fins, rend plus difficile l'obtention d'un résultat proche de la réalité hydraulique du réseau. Augmenter le niveau d'instrumentation en TLRV des compteurs, permettrait un suivi plus fin de ces volumes. Mais l'équipement de la totalité des compteurs est bien trop coûteuse. Il est donc important d'être en mesure d'estimer au mieux les consommations sur une échelle spatio-temporelle fine. Ainsi cette meilleure estimation des consommations permettrait l'obtention de flux hydrauliques plus représentatifs sur le réseau, et également d'améliorer la détection et la spatialisation des fuites.

Chapitre 4

Reconstruction d'un signal partiellement observé sur un graphe par méthode statistique de régression à noyau. Application aux réseaux d'eau potable

Ce chapitre aborde le problème de l'inférence sur les nœuds d'un très grand graphe, représentant un réseau de distribution d'eau potable, à partir d'une observation partielle de quelques données, possiblement chronologiques, sur un faible nombre de nœuds. Nous utilisons une approche de prédiction par noyau reposant sur un estimateur pénalisé de type Ridge qui soulève des problèmes d'analyse spectrale d'une très grande matrice creuse.

Le cadre de prédiction statistique est ici relativement standard puisque qu'on suppose que la structure du graphe est fixée et connue, n'évolue pas avec le temps si on se place dans un aspect dynamique. Seules sont variables les observations effectuées sur les nœuds du graphe. De nombreux travaux traitent de la régression sur données de graphes. Cependant, les verrous existent, tant du point de vue théorique, que pratique :

— L'application de la régression ridge aux données de graphes telle que décrite dans Kolazyk([Kolaczyk, 2009], [Kolaczyk and Csardi, 2014]) mais aussi dans ([Scholkopf and Smola, 2001]) requiert de l'algèbre linéaire sur des matrices carrés de taille le nombre de nœuds. Le stockage des données et le calcul matriciel doivent être appréhendés dans une optique de données massives et d'optimisation des temps de calculs. Il est bien connu que résoudre numériquement un système linéaire ne se fait pas en calculant une inverse de matrices, beaucoup d'algorithmes étant optimisé pour cela ([Livne and Brandt, 2012], [Golub and van Loan, 2013], [Spielman and Teng, 2004]). Ici, et notamment avec la méthode du représentant présentée dans la section 4.1.3, l'effort est pourtant fait une seule fois et peut être payant car il permet par la suite d'accéder au noyau $K^{(n)}$ (de faible dimension) et sa diagonalisation offre une aide à l'interprétation des résultats grâce à la Hat matrix $H(\lambda)$. Par exemple, le Laplacien du graphe pour un cas d'étude comme la ville de Bordeaux est de taille 60 000 x 60 000 et l'application directe de la régression ridge à noyau demande alors une inversion de cette matrice, voire une recherche de son spectre.

— Pire, la reconstruction temporelle des signaux statistiques fait exploser la dimension du problème : si le signal discrétisé en temps est de dimension M, observé sur n_{obs} nœuds parmi N_v nœuds, l'application standard des méthodes à noyau à valeurs vectorielles demande de manipuler un noyau de taille $M.N_v$. Le simple stockage d'une matrice Laplacienne (pleine) demande alors des capacités inaccessibles pour un réseau de l'ordre de la ville de Bordeaux avec un signal mesuré toutes les 5mn sur 2 jours car celle-ci fait 7 953 Téraoctets. Des solutions simplifiées sont donc requises.

— Pourtant, on se place dans un cadre pratique où le ratio entre nœuds observés et nœuds à prédire est en défaveur de la méthode de régression (peu d'observations par rapport à la taille du graphe) et le lissage peut ne refléter qu'une moyennisation du signal sur tous les nœuds non observés. En ce sens le choix du paramètre de lissage en régression ridge est crucial. Les apports théoriques des méthodes de validation croisée ou de validation croisée généralisée doivent être mis en balance de leur applicabilité sur des grands graphes.

— La recherche de solutions par minimisation d'un critère pénalisé (ici en régression ridge) peut être obtenue sans écriture explicite (mais coûteuse) de la solution mais par une méthode itérative de type descente de gradient. Cette méthode évite d'inverser le Laplacien, de stocker de très grosses matrices pleines (non creuses), mais demande d'être validée empiriquement pour avoir une confiance raisonnable dans les résultats de convergence. En outre, les paramètres de "tuning" doivent être judicieusement choisis.

Ainsi, après avoir situé la problématique, nous aborderons dans la suite de ce chapitre les solutions mises en oeuvre pour tenter de palier à ces problèmes.

Nous présentons dans un premier temps les notations ainsi que les méthodes, tirées en grande partie de [Kolaczyk and Csardi, 2014], nous permettant de définir un noyau à partir du graphe représentant le réseau de distribution d'eau potable. Ensuite, nous expliciterons plusieurs formulations des résultats qui peuvent être accessibles avec ou sans diagonalisation ou inversion de grandes matrices. L'explicitation standard du prédicteur sur l'ensemble du graphe ne requiert qu'une inversion de matrice. L'explicitation basée sur un argument RKHS demande une inversion d'une grande matrice ainsi qu'une diagonalisation d'une petite matrice, mais donne toutes les aides aux diagnostiques et la possibilité de validation croisée généralisée. L'utilisation d'une descente de gradient est la méthode la plus rapide, mais ne permet pas le choix du paramètre de lissage.

Enfin, nous présenterons quelques extensions aux signaux temporels ainsi que des résultats obtenus sur les graphes d'Ambès et de Bordeaux, tout d'abord sur données simulées (pour valider la méthode) puis sur données réelles.

4.1 Rappel des méthodes de régression pénalisée sur données de graphes

4.1.1 Les données, les notations

Soit $G = (V, E)$ un graphe non dirigé [*]. G est composé de N_E arrêtes et N_v nœuds [†]. Les nœuds sont identifiés à leur index $i = 1, \dots N_v$. Ainsi deux nœuds i et j sont connectés par une arrête $\{i, j\}$ si une canalisation les relie. On rappelle quelques notations du chapitre deux. La matrice d'adjacence A de G est la matrice (symétrique) de taille $N_v \times N_v$ définie par

$$A_{ij} = \begin{cases} 1, & \text{si} \{i, j\} \in E \\ 0, & \text{sinon} \end{cases}$$

On peut également définir une matrice d'adjacence valuée de la forme :

$$A_{ij} = \begin{cases} w_{ij} > 0, & \text{si} \{i, j\} \in E \\ 0, & \text{sinon} \end{cases}$$

où $w_{ij} = w_{ji}$ mesure une force de connexion entre deux nœuds i et j.

On définit également la matrice des degrés du graphe G $d_i = \sum_{j \neq i} Aij$ comme $\mathbf{D} = \text{diag}[(d_i)_{i \in V}]$ et enfin le Laplacien de G, $\mathbf{L} = \mathbf{D} - \mathbf{A}$.

Soit $X = (X_1, ..., X_{N_v})$ les attributs des noeuds du graphe G. On suppose observer les valeurs de $X_i = x_i$ pour $i \in V^{obs} \subseteq V$ et stocker dans le vecteur $x_{obs} = (x_i)_{i \in V_{obs}}$. On note alors $n = |V^{obs}|$ le nombre de nœuds sur lesquels le processus X est observé (par exemple les nœuds représentant les appareils de mesure). Pour simplifier l'exposé, on suppose dans un premier temps que $X \in \mathbb{R}$. Notre objectif est de construire un prédicteur $\hat{h}$ de V dans $\mathbb{R}$ permettant de prédire les valeurs de X sur les nœuds non observés tout en s'ajustant au mieux aux données observées.

Dans la suite de ce chapitre, on considère un prédicteur basé sur une régression pénalisée de type ridge, dont la pénalisation, ainsi que l'ensemble des régresseurs, vont être adaptés à la structure du graphe. Cette adaptation utilise les méthodes à noyau pour définir une solution des moindres carrés régularisée appelée *Kernel ridge regression* (KRR).

La fonction $\mathbf{f}(.)$ sur V peut être représentée comme un vecteur $\mathbf{f} \in \mathbb{R}^{N_v}$ avec le i-ème élément de $\mathbf{f}$ associé au nœud $i \in V$. Considérant les observations $\mathbf{x}^{obs}$ et le graphe G, l'objectif est ici de construire un prédicteur linéaire selon un ensemble de régresseurs à choisir et adaptés au graphe, avec un compromis d'attache aux données sur les nœuds observés et de lissage de la fonction h sur l'ensemble des nœuds non observés. Le cadre de la régression pénalisée permet ceci en définissant l'ensemble des régresseurs $(\phi_1, \dots \phi_{N_\nu})$ et une pénalisation $pen(\beta)$ en vue de la minimisation d'un critère du type :

$$\sum_{i \in V^{obs}} (x_i - (\Phi\beta)_i)^2 + \lambda pen(\beta) \tag{4.1.1}$$

∗. représentant par exemple le réseau de distribution d'eau potable de la métropole bordelaise. Les arrêtes E (resp. nœuds V) de G symbolisent les canalisations (resp. les objets sur lesquels les canalisations sont connectées, tels que des raccords, des vannes, des appareils de mesures etc.)

†. pour rappel sur Bordeaux $N_E = 66\ 482$ arrêtes et $N_v = 60\ 097$ nœuds

pour λ un réel positif. La régression ridge à noyau adaptée au cadre d'un graphe, ici connecté et non dirigé, justifie les choix particuliers pour les régresseurs et ainsi que la pénalité, comme décrit ci-dessous.

Avant de définir l'estimateur ridge choisi, on rappelle quelques propriétés utiles du Laplacien vues au Chapitre 2. Si $L = D - A$,

1. L est diagonalisable et on note ses valeurs propres par ordre croissant $\lambda_1 = 0 < \lambda_2 \leq \cdots \leq \lambda_{N_v}$
2. L admet la diagonalisation $L = \Phi \Delta_L \Phi^T = \sum_{i=1}^{N_v} \lambda_i \phi_i \phi_i^T$ où la matrice diagonale contient les valeurs propres $\lambda_i, i = 1 \dots N_v$ et $\Phi = [\phi_1, \dots, \phi_{N_v}]$ est la matrice orthogonale des vecteurs propres normés : $\Phi^T \Phi = Id_{\mathbb{R}^{N_v}}$.
3. Le vecteur propre ϕ_1 associé à la seule valeur propre nulle est constant : $\phi_1 = 1/\sqrt{N_v}\mathbf{1}$.
4. De l'item précédent, il découle que les vecteurs propres suivants $\phi_2, \dots \phi_{N_v}$ sont centrés, car orthogonaux à ϕ_1.
5. L'image de l'application L est ainsi formée des vecteurs centrés de $\mathbb{R}^{N_v}$ et $\forall f \in \mathbb{R}^{N_v}, f^T L f = \sum_{i,j \in V_{obs}} w_{ij}(f_i - f_j)$

Soit $K = L^+$ la pseudo inverse de Moore Penrose [‡] du Laplacien L du graphe définie par $\mathbf{K = \Phi \Delta \Phi^T}$, $\mathbf{\Delta}$ étant la matrice diagonale des valeurs propres de K par ordre croissant (sauf la première, qui est nulle) : $0 < \delta_2 = 1/\lambda_2 \geq \cdots \geq 1/\lambda_{N_v}$.

$$\Delta = diag(0, \delta_i = 1/\lambda_i, i \geq 2) = \begin{pmatrix} 0 & & & & \\ & \delta_1 = \frac{1}{\lambda_1} & & & \\ & & \delta_2 = \frac{1}{\lambda_2} & & \\ & & & .. & \\ & & & & \delta_n = \frac{1}{\lambda_{N_v}} \end{pmatrix}$$

Ainsi on a :

$$K = L^+ = \Phi \Delta \Phi^T, \quad L = K^+ = \Phi \Delta_L \Phi^T, \quad \Phi \Phi^T = Id$$

Bien sûr $\Delta^+ = \Delta_L$, matrice diagonale des valeurs propres du Laplacien.

4.1.2 Régression ridge pour un noyau Laplacien

En suivant Kolaczyk ([Kolaczyk and Csardi, 2014], [Kolaczyk, 2009]), on définit l'estimateur ridge sur un graphe ci-dessous.

Définition 4.1.1. Estimateur ridge sur le graphe de Laplacien L pour une fonction réelle

Soit un graphe connecté non dirigé $G = (V, E)$ et son Laplacien $L = D - A$ avec les notations définies ci-dessus, A étant une matrice d'adjacence possiblement valuée. Soit $x = (x_1, \dots, x_{N_v})$ les valeurs d'une fonction réelle $\mathbf{f}$ sur V et $x_{obs} = (x_i, i \in V_{obs}) \in \mathbb{R}^n$

‡. La pseudo-inverse de Moore Penrose d'une matrice M carrée (ici symétrique) est une matrice M^+ vérifiant $M^+MM^+ = M^+$ et $MM^+M = M$, existant même si M n'est pas inversible. L'écriture de ce pseudo-inverse de Moore Penrose selon les vecteurs propres de M diagonalisable est : $M^+ = \Phi \Delta^{-1} \Phi$, où la matrice diagonale Δ^{-1} contient les inverses des valeurs propres non nulles de M, et zéro ailleurs.

les valeurs observées sur les nœuds V_{obs}. Soit Δ la matrice diagonale des valeurs propres de $K = L^+$ et $\Delta^+ = \Delta_L$ sa pseudo inverse. Pour la base de régresseurs $(\phi_i)_{i=1\dots N_v}$ et le prédicteur associé $h = \Phi\beta$, l'estimateur ridge $\hat{\beta}$ de la régression est défini par

$$\hat{\beta} = \underset{\beta \in \mathbf{R}^{N_\nu}}{\arg\min} \sum_{i \in V^{obs}} (x_i - (\phi\beta)_i)^2 + \lambda \beta^T \Delta_L \beta \tag{4.1.2}$$

Le prédicteur associé est $\hat{f}$ défini par

$$\hat{f} = \mathbf{\Phi}\hat{\beta}$$

Cette définition d'une régression pénalisée d'un vecteur de coefficient β associée à un ensemble bien choisi de variables $\phi_1, \dots, \phi_{N_v}$ ainsi qu'une pénalisation pondérée par la matrice diagonale des valeurs propres est assez classique. Ici il faut noter que le nombre de covariables vaut exactement le nombre total de nœuds dans le graphe, bien supérieur au nombre d'observations. Nous justifierons dans la section suivante la cohérence de cette définition par un argument RKHS (espace fonctionnel de Hilbert à noyau reproduisant), mais quelques remarques s'imposent déjà, qui permettent une ré-écriture explicite du prédicteur et de l'estimateur $\hat{\beta}$:

Remarque 1 : ré-écriture de la pénalisation

Si on s'intéresse au terme de pénalisation dans (4.1.2) on remarque que

$$\begin{aligned} \beta^T \Delta_L \beta &= \beta^T \Phi^T \Phi \Delta_L \Phi^T \Phi \beta \\ &= f^T L f, \end{aligned}$$

d'une part, et que

$$f^T L f = \sum_{i,j \in E} w_{ij} (f_i - f_j)^2 \tag{4.1.3}$$

d'autre part. Ainsi, choisir la base de fonctions $(\phi_k)_{k=1\dots N_v}$ et la pénalisation $\beta^T \Delta_L \beta$ permet une régression adaptée à la structure du graphe : on cherche à lisser les valeurs de $\mathbf{f}$ assignées aux nœuds i et j adjacents dans G pour une fonction dans l'espace linéaire des vecteurs propres de L.

Remarque 2 : Interprétation en fréquence des coefficients β

[Romero et al., 2017] ont une interprétation intéressante de la décomposition du signal f en terme de transformée de Fourier sur un graphe : les $\phi_k, k = 1 \dots N_v$, vecteurs propres du Laplacien peuvent être assimilés à une base de Fourier car ils décomposent f en N_v coefficients $\beta_1, \dots \beta_{N_v}$ associés à des "fonctions" dont les normes (les énergies) sont croissantes : $\|\phi_k\|^2 = \Phi_k^T L \Phi_k = \lambda_k$. Ainsi, les vecteurs propres de L associés aux plus petites valeurs propres reflètent des fonctions régulières sur le graphe tandis que les Φ_k associés aux grandes valeurs propres λ_k sont des fonctions à haute fréquence. Cet aspect a déjà été développé dans la Section 2.4 du Chapitre 2 pour le vecteur de Fiedler et la représentation des "bas" vecteurs propres sur le graphe.

Puisque $f = \sum_{k=1}^{N_v} \beta_k \Phi_k$, la pénalisation $f^T L f$ s'écrit donc aussi

$$f^T L f = \sum_{i,j \in E} w_{ij} (f_i - f_j)^2 = \beta^T \Delta_L \beta = \sum_{k=1}^{N_v} \lambda_k \beta_k^2$$

La pénalisation donne un plus grand poids λ_k dans la pénalité associée au coefficient β_k.

Remarque 3 : La pénalisation $\beta^T \Delta_L \beta$ n'est qu'un des exemples possibles et Kolaczyk ([Kolaczyk, 2009], Belkin et al [Belkin et al., 2004] proposent d'étendre à des pénalisations basées sur une transformation du Laplacien. Pour r une fonction bien choisie, par exemple, $r(x) = x^p$, $r(x) = exp(ux)$, les extensions possibles sont $pen(\beta) = \beta^T r(\Delta)\beta$, ce qui revient à utiliser la matrice *de lissage* L^p ou $\exp(uL)$ au lieu du Laplacien de graphe dans l'expression $f^T L f$.

Remarque 4 : Nous imposerons la condition de centrage des candidats $f = \Phi\beta$ afin de garantir que l'espace des fonctions dans lequel on cherchera un minimum a la propriété d'être un espace fonctionnel avec une structure hilbertienne bien définie. En particulier, les fonctions constantes devront être nécessairement nulles, ce qui veut dire que $\beta_1 = 0$ et f sera une combinaison linéaire des vecteurs propres (centrés) $\phi_2, \ldots, \phi_{N_v}$. (On rappelle que zero est une valeur propre de L puisque $L.\mathbf{1} = 0$. $\mathbf{1}$ est donc vecteur propre de L associé à la valeur propre nulle et tous les autres vecteurs propres ϕ_k seront orthogonaux à $\mathbf{1}$).

Dans la pratique, on centrera donc les données observées $z_{obs} = x_{obs} - \mu_{obs}$ avant de chercher un prédicteur $\hat{f}$, puis les prédictions seront obtenues par $\hat{f} + \mu_{obs}$.

4.1.2.a Explicitation standard de l'estimateur

Dans cette section, on détermine une forme explicite pour l'estimateur $\hat{\beta}$ ainsi que le prédicteur $\hat{f}$.

Soit le masque $\mathbf{\Omega}$, matrice de taille $n \times N_v$ dont les entrées sont : $\Omega_{ij} = 1$ si $j \in V^{obs}$, et zéro sinon, pour tout $i \in V^{obs}$. Ainsi, on peut passer du vecteur total des données $z \in \mathbb{R}^{N_v}$ au vecteur des données observées $z_{obs} = (x_i, i \in V_{obs})$ par le produit $z_{obs} = \Omega z$:

$$z_{obs} = \begin{pmatrix} z_{i_1} \\ . \\ . \\ z_{i_n} \end{pmatrix} = \Omega\ z = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} z_1 \\ . \\ . \\ . \\ . \\ . \\ z_{N_v} \end{pmatrix}$$

L'estimateur $\hat{\beta}$ solution du problème d'optimisation vérifie donc

$$\hat{\beta} \quad = \quad \underset{\beta \in \mathbb{R}^{N_v}}{\arg\min} \sum_{i \in V^{obs}} (z_i - (\Phi\beta)_i)^2 + \lambda \beta^T \Delta_L \beta$$

qui est un problème équivalent à déterminer $\hat{f} = \Phi\hat{\beta}$ avec

$$\begin{aligned} \hat{f} &= \underset{f \in \mathbb{R}^{N_v}}{\arg\min} \sum_{i \in V^{obs}} (z_i - f_i)^2 + \lambda f^T L f \\ &= \underset{f \in \mathbb{R}^{N_v}}{\arg\min} \|\mathbf{\Omega} z - \mathbf{\Omega} f\|^2_{\mathbb{R}^n} + \lambda f^T L f \\ &= z^T M z - 2 f^T M z + f^T (M + \lambda L) f, \end{aligned} \tag{4.1.4}$$

où $M = \mathbf{\Omega}^T\mathbf{\Omega}$ est la matrice diagonale de taille N_v telle que $M_{ii} = 1$ si $i \in V_{obs}$. Notons que $z \in \mathbb{R}^{N_v}$ contient les vraies valeurs centrées (par μ_{obs}), dont seules sont observées les coordonnées d'indice $i \in V_{obs}$. Ainsi, $Mz \in \mathbb{R}^{N_v}$ prend des zéros pour valeurs sauf aux indices observés. On replonge ainsi z_{obs} dans $\mathbb{R}^{N_v}$ et on peut noter $Mz = z_{obs}^{N_v} \in \mathbb{R}^{N_v}$.

Par dérivation matricielle, on montre donc que le minimum est atteint pour f solution de

$$(M + \lambda L)\, f = z_{obs}^{N_v} \tag{4.1.5}$$

Notons que le Laplacien n'est pas inversible, mais qu'une fois sa diagonale perturbée par M (ajout d'une constante positive **uniquement** sur les coordonnées des points observées), on obtient une matrice $(M + \lambda L)$ qui devient inversible car elle est à diagonale fortement dominante et irréductible [§]. En effet, $|(M+\lambda L)_{ii}| = 1_{i\in V_{obs}} + \lambda|L_{ii}|$ est supérieur (ou égal) à la somme des valeurs absolues des autres coefficients de la ligne i, c'est à dire $\lambda|D_{ii}| = \lambda|L_{ii}|$. D'après Horn et Johnson, 2013 ([Horn and Johnson, 2012], Th. 6.1.11, p. 392), cela implique l'inversibilité. L'estimateur pénalisé est donc bien défini, on obtient :

$$\hat{f} = (M + \lambda L)^{-1} Mx = (M + \lambda L))^{-1} z_{obs}^{N_v} \tag{4.1.6}$$

et

$$\hat{\beta} = \Phi^T \hat{f} = \Phi^T (M + \lambda L))^{-1} z_{obs}^{N_v}. \tag{4.1.7}$$

Il faut noter que β n'est que la représentation de la fonction f dans la base des vecteurs propres de L, qui en tant que régresseurs, engendrent $\mathbb{R}^{N_v}$ tout entier. On verra que l'optimum $\hat{f}$ est nécessairement centré.

On peut noter aussi que les vecteurs propres convenablement ordonnés définissent en fait une base de Fourier et les coefficients dans le vecteur $\hat{\beta}$ sont assimilés à des coefficients de Fourier pour cette base $\phi_2, \dots \phi_n$. L'estimation de ces coefficients cherche à pénaliser les coefficients associés aux vecteurs propres à "haute fréquence" c'est à dire ceux qui ne sont pas "lisses" sur le graphe.

Cette expression requière l'inversion d'une matrice de taille $N_v \times N_v$, ce qui peut être prohibitif pour des ordinateurs standards, et même des serveurs de calcul. A titre d'exemple, le stockage seul de la matrice associée au graphe du réseau de Bordeaux (57 400 nœuds), demande une capacité de 24,5 Go. De nombreux algorithmes résolvent l'équation $Af = x$ sans inverser la matrice A. Nous présentons ci-dessous un résultat par descente de gradient.

4.1.2.b Descente de gradient pour résoudre $(M + \lambda L)f = z_{obs}^{N_v}$

On cherche à résoudre l'équation (4.1.5) sans avoir à inverser la matrice carrée symétrique définie positive $A = (M + \lambda L)$ de taille $(N_v \times N_v)$. Le gradient de la fonctionnelle à minimiser $\mathcal{F}(f) = \|\mathbf{\Omega}z - \mathbf{\Omega}f\|^2_{\mathbb{R}^n} + \lambda f^T L f$ est

$$\bigtriangledown(f) = 2(M + \lambda L)f - 2\Omega^T z_{obs}^{N_v} = 2Af - 2b$$

§. Un graphe G est connecté si et seulement si sa matrice $L(G)$ est irréductible, de même que $M + \lambda L$

où $A = (M + \lambda L)$ matrice carrée symétrique définie positive de taille $(N_v \times N_v)$, $M = \Omega^T \Omega$ la matrice diagonale de taille N_v telle que $M_{ii} = 1$ si $i \in V_{obs}$ et $b = \Omega^T z$, avec $z \in \mathbb{R}^{N_v}$ contient les vraies valeurs centrées (par μ_{obs}), dont seules sont observées les coordonnées d'indice $i \in V_{obs}$.

Le minimum de cette fonctionnelle est atteint pour le point $\mathcal{F}(f^*) \in \mathbf{R}^{N_v}$ qui annule $\bigtriangledown(f)$

$$\bigtriangledown(f^*) = 0 \iff Af^* = b$$

La méthode du gradient conjugué fait partie des méthodes de descente qui ont comme principe commun la recherche de f suivant l'algorithme itératif

$$f_0 \text{ donné, } f_{k+1} = f_k + \alpha_k p_k, \ \forall\ 0 \leq k \leq n$$

avec $p_k \in \mathbf{R}^{N_v}$ la direction de la descente et $\alpha_k \in \mathbf{R}$ le pas de la descente. Les directions de la descente p_k sont construites de telle manière que $r_k = Ax_k - b$ soient tous orthogonaux entre eux. i.e. $\langle r_i, r_j \rangle = 0, \forall\ 0 \leq i \leq j \leq n$

Théorème 4.1.1. *L'algorithme de descente de gradient conjugué s'arrête en au plus n étapes et on a pour tout k, tant que l'algorithme n'a pas terminé (c'est à dire si $r_k \neq 0$), on a l'égalité entre les sous-espaces suivants :*

$$(r_0, Ar_0, \ldots, A^k r_0) = (r_0, r_1, \ldots, r_k) = (p_0, p_1, \ldots, p_k)$$

De plus, pour tout i tel que $0 \leq i \leq k$, on a les relations d'orthogonalité suivantes :
(i) $\langle p_{k+1}, Ap_i \rangle = 0$,
(ii) $\langle r_{k+1}, p_i \rangle = 0$,
(iii) $\langle r_{k+1}, r_i \rangle = 0$

Pour montrer que l'algorithme s'arrête en au plus n étapes, il suffit de remarquer que tant que $r_k \neq 0$, la famille $(r_i)_{0 \leq i \leq k}$ est donc une famille orthogonale et donc ne peut avoir qu'au plus n vecteurs non-nuls, en particulier on a donc $r_n = 0$.

Ainsi les directions de descente sont toutes conjuguées deux à deux pour la matrice A, et les gradients r_k sont tous orthogonaux deux à deux pour le produit scalaire usuel. Par conséquents, ces familles forment des familles orthogonales et ne peuvent être non-nulles qu'un nombre de fois limité par la dimension de l'espace. C'est pourquoi l'algorithme s'arrête en moins de n étapes.

Algorithme 3 Descente de gradient conjugué

Entrée:

$A = (M + \lambda L)$ *avec* $M = \Omega^T \Omega$

$b = \Omega^T z$

$r_0 := b - Ax_0$

$p_0 := r_0$

$k := 0$

Sortie: x_{k+1}

tant que $\| r_k \| > \epsilon$ **faire**

$\alpha_k := \frac{r_k^T r_k}{p_k^T A p_k}$ *(pas optimal)*

$x_{k+1} := x_k + \alpha_k p_k$ *(avancement suivant)*

$r_{k+1} := r_k - \alpha_k A p_k$ *(calcul itératif du gradient)*

$\beta_k := \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$ *(paramètre assurant* $< x_{k+1}, x_k >= 0$*)*

$p_{k+1} := r_{k+1} + \beta_k p_k$ *(calcul itératif de la direction de la descente)*

$k := k + 1$

fin tant que

4.1.3 Interprétation RKHS de la régression pénalisée - théorème du représentant

Dans cette section, nous utilisons le formalisme de l'optimisation dans un espace de Hilbert à noyau reproduisant ([Hastie et al., 2009], [Romero et al., 2017]) pour montrer que le problème cité admet une solution dans un espace de dimension réduite par rapport à l'espace initial ($\mathbb{R}^{N_v}$) dans lequel on cherche une solution. On se restreint toujours à l'approximation d'une fonction réelle $f : G \to R$, en notant x le vecteur de $\mathbb{R}^{N_v}$ de ses valeurs.

4.1.3.a Cas général d'une estimation régularisée dans un RKHS de dimension finie

Le problème d'optimisation pénalisée (4.1.5) est ici interprété comme une optimisation régularisée dans un espace fonctionnel. Nous construisons l'espace de fonctions à partir d'un noyau défini positif K, défini ci-dessous.

Définition 4.1.2. K est un noyau défini positif sur l'ensemble V si et seulement si c'est une fonction :$K : V \times V \to \mathbb{R}$, symétrique, et vérifiant

$$\forall n \in \mathbb{N}, (v_1, \ldots, v_N) \in V^{\mathbb{N}}, (a_1, \ldots, a_n) \in \mathbb{R}^N, \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K(v_i, v_j) \geq 0.$$

En d'autres termes, toutes les matrices de Gram $\boldsymbol{K} = [K]_{ij}$ d'évaluation du noyau sur l'ensemble quelconque $(v_1, \ldots, v_N) \in V^N$ sont des matrices symétriques définies positives.

On interprétera les valeurs de K comme des mesures (positives) de similarité entre les points v de V (les nœuds du graphe).

Le noyau défini positif K permet de créer un unique espace de Hilbert de fonctions de V dans $\mathbb{R}^{N_v}$. On rappelle pour cela le théorème fondamental (Aronzaj, 1950) :

Théorème 4.1.2. *K est un noyau défini positif sur l'ensemble $\mathcal{X}$ si et seulement si il existe un espace de Hilbert $\mathcal{H}$ et une application $\Psi : \mathcal{X} \to \mathcal{H}$, telle que,*

$$\forall x, x' \in \mathcal{X} : K(x, x') =< \Psi(x), \Psi(x') >_{\mathcal{H}}$$

Cet espace de Hilbert est appelé Espace de Hilbert à noyau reproduisant (Reproducing Kernel Hilbert space) car :

1. *il contient toutes les fonctions $K_x : t \to K(x, t)$, pour $x \in \mathcal{X}$,*
2. *il bénéficie de la propriété autoreproduisante : $\forall x \in \mathcal{X}$ et $f \in \mathcal{H}$, $f(x) =< f, K_x >_{\mathcal{H}}$.*

En dimension finie, c'est ici assez facile d'expliciter la construction de l'espace $\mathcal{H}$ à partir du noyau K, et de justifier le choix du noyau K pour des considérations d'approximations de fonctions "lisses" sur le graphe G. En fait, l'usage des $RKHS$ est justifié par les résultats induits en théorie de l'apprentissage dans des ensembles quelconques, et notamment le théorème du représentant, qui nous permettra dans notre problème d'obtenir une expression utile de la solution de régression ridge sur le graphe. Adapté à nos notations, nous donnons le

Théorème 4.1.3. *Théorème du représentant ([Kimeldorf and Wahba, 1971], [Scholkopf and Smola, 2001])*

Soit $\mathcal{H}$ un RKHS de fonctions de V dans $\mathbb{R}$: $v \in V \to f(v) \in \mathbb{R}$, et l'ensemble des observations $(i, x_i))_{i \in V_{obs}}$ sur un sous-ensemble de nœuds du graphe. Alors la solution de

$$min_{f \in \mathcal{H}} \sum_{i \in V_{obs}} (x_i - f(i))^2 + \lambda \|f\|^2_{\mathcal{H}} \tag{4.1.8}$$

admet une représentation de la forme :

$$\forall v \in V, f(v) = \sum_{i=1}^{n} \alpha_i K(i, v),$$

codée vectoriellement

$$f = \boldsymbol{K}^{N_v, n} \alpha$$

où $\alpha \in \mathbb{R}^n$ et $\boldsymbol{K}^{(N_v, n)}$ est la matrice de taille $N_v \times n$ qui contient les n colonnes de $\boldsymbol{K}$ associées aux nœuds observés x_{obs}.

Ce résultat nous dit simplement qu'au lieu de chercher la solution dans l'espace entier $\mathcal{H}$ (ici de dimension finie au plus N_v), il suffit de le chercher dans l'espace linéaire engendré par l'ensemble des n fonctions $K(v_i, .), v_i \in V_{obs}$. La dimension du problème en est donc réduite.

En outre, puisque le résultat minimisant la fonctionnelle est une combinaison linéaire des colonnes de $\boldsymbol{K}$, celui-ci est nécessairement centré car toutes les colonnes de $\boldsymbol{K}$ sont

centrées. En effet,

$$
\begin{aligned}
1^T \boldsymbol{K} &= 1^T \sum_{i=1}^{N_v} \delta_i \phi_i \phi_i^T && (4.1.9)\\
&= \sum_{i=1}^{N_v} \delta_i (1^T \Phi_i) \Phi_i^T && (4.1.10)\\
&= \sum_{i=1}^{N_v} \delta_i < 1, \Phi_i >_{\mathbb{R}^{N_v}} \Phi_i^T && (4.1.11)\\
&= 0 && (4.1.12)
\end{aligned}
$$

car $\delta_1 = 0$ et $< 1, \Phi_i >_{\mathbb{R}^{N_v}} = 0$ pour $i = 2, \dots N_v$.

De plus, ce théorème nous permet donc d'écrire que, à partir de $f = \boldsymbol{K}^{N_v,n}\alpha$, et en notant $\boldsymbol{K}^{(n)}$ la matrice (carrée) de Gram $\boldsymbol{K}^{(n)} = K(i,j), i,j \in V_{obs}$, on a

$$\hat{f}_{obs} = (\hat{f}(v_1), \dots, \hat{f}(v_n))^T = \boldsymbol{K}^{(n)}\alpha$$

donc

$$
\begin{aligned}
\|\hat{f}\|_{\mathcal{H}} = \|\sum_{i=1}^{n} \alpha_i K(v_i, .)\|_{\mathcal{H}}^2 &= \|\sum_{i=1}^{n} \alpha_i K_{v_i}\|_{\mathcal{H}}^2 && (4.1.13)\\
&= \sum_{i,j \in V_{obs}} \alpha_i \alpha_j < K_{v_i}, K_{v_j} >_{\mathcal{H}} && (4.1.14)\\
&= \sum_{i,j \in V_{obs}} \alpha_i \alpha_j \boldsymbol{K}_{ij} && (4.1.15)\\
&= \alpha^T \boldsymbol{K}^{(n)} \alpha && (4.1.16)
\end{aligned}
$$

et

$$\sum_{i \in V_{obs}} (z_i - f(i))^2 = \sum_{i \in V_{obs}} (z_i - (\boldsymbol{K}^{(n)}\alpha)_i)^2 = \|z_{obs} - \boldsymbol{K}^{(n)}\alpha\|^2$$

Le problème d'optimisation est donc réduit à

$$\arg\min_{\alpha \in \mathbb{R}^n} \|z_{obs} - \boldsymbol{K}^{(n)}\alpha\|_{\mathbb{R}^n}^2 + \lambda \alpha^T \boldsymbol{K}^{(n)} \alpha$$

dont la solution est, après dérivation matricielle, et puisque $\boldsymbol{K}^{(n)} + \lambda I$ est inversible,

$$\hat{\alpha} = (\boldsymbol{K}^{(n)} + \lambda I)^{-1} z_{obs}. \quad (4.1.17)$$

Nous avons donc une nouvelle formulation pour le prédicteur $\hat{f}$:

$$\hat{f} = \boldsymbol{K}^{N_v,n}\hat{\alpha} = \boldsymbol{K}^{N_v,n}(\boldsymbol{K}^{(n)} + \lambda I)^{-1} z_{obs}, \quad (4.1.18)$$

soit

$$\hat{f} = \boldsymbol{K}^{N_v,n} \Phi^{(n)} diag\left(\frac{1}{\delta_i^{(n)} + \lambda}\right) \Phi^{(n)T} z_{obs}, \quad (4.1.19)$$

où $\Phi^{(n)}$ est la matrice des vecteurs propres de $\boldsymbol{K}^{(n)}$ associés aux valeurs propres $\delta_1^{(n)} \geq \ldots \delta_n^{(n)} \geq 0$.

En chaque nœud $i \in V$, la prédiction est donc obtenue comme combinaison linéaire des fonctions $K_j, j \in V_{obs}$ dont les coefficients sont dans $\hat{\alpha} = (\boldsymbol{K}^{(n)} + \lambda I)^{-1} z_{obs}$. Le vecteur $\hat{\alpha}$ nous renseigne sur l'importance globale de chaque nœud observé i dans la reconstruction du signal sur le graphe entier.

Enfin, une autre interprétation est utile : soit

$$H = \boldsymbol{K}^{N_v,n}(\boldsymbol{K}^{(n)} + \lambda I)^{-1},$$

la *Hat matrix* globale, on a

$$\hat{f} = H z_{obs}.$$

Les coefficients de la *Hat matrix* nous renseignent sur la façon dont chaque valeur $f_i, i \in V$ est reconstruite comme combinaison particulière des observations z_{obs}. La i-ème ligne de H contient les coefficients à appliquer à z_{obs} pour calculer $\hat{f}_i$. En particulier, concernant l'attache aux données, en se restreignant aux valeurs observées, on a

$$\begin{aligned} \hat{f}_{i,i\in V_{obs}} &= \boldsymbol{K}^{(n)}(\boldsymbol{K}^{(n)} + \lambda I)^{-1} z_{obs} && (4.1.20) \\ &= \Phi^{(n)} diag\left(\frac{\delta_i^{(n)}}{\delta_i^{(n)} + \lambda}\right) \Phi^{(n)T} z_{obs} && (4.1.21) \end{aligned}$$

En particulier, on peut en déduire une expression basée sur les valeurs propres pour la somme des carrés des résidus du critère de minimisation :

$$\|z_{obs} - \boldsymbol{K}^{(n)}\hat{\alpha}\|^2_{\mathbb{R}^n} = \|z_{obs} - \hat{f}_{i,i\in V_{obs}}\|^2_{\mathbb{R}^n} = z_{obs}^T \Phi^{(n)} diag\left(\frac{\lambda^2}{(\delta_i^{(n)} + \lambda)^2}\right) \Phi^{(n)T} z_{obs}/$$

4.1.3.b Application au noyau dérivé du Laplacien de graphe

Dans notre cadre, il est possible d'expliciter un peu plus le noyau et le produit scalaire du RKHS. Calculons d'abord la fonction "mapping" $\psi : V \to \mathbb{R}^{N_v}$. Puisque $\boldsymbol{K} = \Phi\Delta\Phi^T$ est le pseudo inverse du Laplacien L, on peut écrire l'application ψ telle que $K(x,x') =< \psi(x), \psi(x') >_{\mathcal{H}}$, ou, noté matriciellement :

$$\boldsymbol{K}_{ij} =< \psi(x_i), \psi(x_j) >_{\mathcal{H}} .$$

En effet,

$$\begin{aligned}
\boldsymbol{K} &= \Phi\Delta\Phi^T = \sum_{i=1}^{N_v} \delta_i\phi_i\phi_i^T \\
&= \sum_{i=1}^{N_v} \sqrt{\delta_i}\phi_i(\sqrt{\delta_i}\phi_i)^T \\
&= \begin{bmatrix} | & & | \\ \sqrt{\delta_1}\phi_1 & & \sqrt{\delta_{N_v}}\phi_{N_v} \\ | & & | \end{bmatrix} \begin{bmatrix} — & \sqrt{\delta_1}\phi_1 & — \\ — & — & — \\ — & \sqrt{\delta_{N_v}}\phi_{N_v} & — \end{bmatrix} \\
&= \begin{bmatrix} — & \psi_1 & — \\ — & — & — \\ — & \psi_{N_v} & — \end{bmatrix} \begin{bmatrix} | & & | \\ \psi_1 & & \psi_{N_v} \\ | & & | \end{bmatrix} \\
&= \Psi^T\Psi = \sum_{i=1}^{N_v} \psi_i^T\psi_i
\end{aligned} \tag{4.1.22}$$

où la matrice Ψ admet en colonne les vecteurs

$$\psi_i = \psi(v_i) = \begin{pmatrix} \sqrt{\delta_1}\phi_1(i) \\ \sqrt{\delta_2}\phi_2(i) \\ . \\ . \\ \sqrt{\delta_{N_v}}\phi_{N_v}(i) \end{pmatrix} \tag{4.1.23}$$

Ainsi, on peut écrire le produit scalaire induit dans la construction du RKHS à partir du noyau $\boldsymbol{K}$:

$$\begin{aligned}
K &= \begin{bmatrix} — & \psi(v_1) & — \\ — & — & — \\ — & \psi(v_n) & — \end{bmatrix} \begin{bmatrix} | & & | \\ \psi(v_1) & & \psi(v_n) \\ | & & | \end{bmatrix} \\
&= [< \psi(v_i), \psi(v_j) >_{\mathbb{R}^n}] \\
&= \left[\psi(v_i)^T\psi(v_j)\right] = \left[K(v_i, v_j)\right],
\end{aligned} \tag{4.1.24}$$

Bien que cela ait peu d'utilité, il est possible d'exprimer $\hat{\beta}$ en fonction de $\hat{\alpha}$. Puisque $\hat{f} = \Phi\hat{\beta}$, et Φ orthogonale, on a

$$\hat{\beta} = \Phi^T\boldsymbol{K}^{N_v,n}\hat{\alpha} = \Phi^T\boldsymbol{K}^{N_v,n}(\boldsymbol{K}^{(n)} + \lambda I)^{-1}z_{obs}.$$

4.1.4 Résumé des différentes formes du prédicteur et coût computationnel

Nous avons vu que la solution de (4.1.8) admet plusieurs expressions explicites. A partir du Laplacien de graphe, facile à calculer, les efforts minimums sont

1. **Méthode1 :** l'inversion directe avec la méthode du masque donne

$$\hat{f} = (M + \lambda L)^{-1} Mx = (M + \lambda L))^{-1} x_{obs}^{N_v}.$$

Nous avons simplement à déterminer la pseudo-inverse d'une matrice symétrique de grande taille pour le calcul de $\hat{f}$. Il faut éventuellement diagonaliser L (ce qui est inaccessible pour de grands graphes) pour déterminer $\hat{\beta} = \Phi^T \hat{f} = \Phi^T \left(M + \lambda L\right))^{-1} x_{obs}^{N_v}$.

2. **Méthode2 :** l'expression avec la méthode du représentant donne $\hat{f} = \boldsymbol{K}^{N_v,n}\hat{\alpha} = \boldsymbol{K}^{N_v,n}(\boldsymbol{K}^{(n)} + \lambda I)^{-1}x_{obs}$, ce qui nécessite la détermination du noyau K, donc le calcul de l'inverse de Moore-Penrose du Laplacien L. Ensuite, selon l'expression utilisée pour $\hat{f}$, on peut avoir à inverser une matrice de taille n ($(\boldsymbol{K}^{(n)} + \lambda I)$, ou à diagonaliser $\boldsymbol{K}^{(n)}$, ce qui peut être raisonnable quand le nombre d'observations est faible par rapport à la taille du graphe.
3. **Méthode indirecte :** la descente de gradient conjugué, demande à chaque itération un nombre réduit d'opérations matricielles : une multiplication d'une matrice (creuse) de taille N_v par un vecteur de $\mathbb{R}^{N_v}$ (pour le calcul du pas optimal) et des manipulations de vecteurs de tailles N_v, le nombre maximum d'itérations étant N_v (rarement atteint en pratique).

Ainsi, il est possible d'éviter la diagonalisation de L pour la simple recherche du prédicteur. Il est toujours intéressant mathématiquement d'exprimer les résultats dans la base des vecteurs propres mais en pratique, seuls quelques vecteurs propres peuvent être obtenus dans un temps raisonnable, ceci même sur des machines multicœurs à fort stockage en mémoire. Nous préciserons plus loin les considérations matérielles sur les cas concerts des graphes d'Ambes (1 825 nœuds) et de Bordeaux (57 400 nœuds).

4.1.5 Inversion du Laplacien

Dans cette section, on présente l'algorithme utilisé pour déterminer la pseudo-inverse de Moore-Penrose du Laplacien, sans avoir recours à la recherche de ses vecteurs propres. Il faut noter qu'une autre méthode a aussi été envisagée : évaluer un sous-ensemble de vecteurs propres de L pour déterminer une approximation de type

$$L = \sum_{i=1}^{N_v} \lambda_i \Phi_i \phi_i^T \sim \tilde{L} = \sum_{i \in I} \lambda_i \Phi_i \phi_i^T$$

où l'ensemble I des vecteurs propres sélectionnés peuvent être ceux associés aux p plus grandes ou p plus petites valeurs propres. En termes d'approximation sur le graphe, pour une fonction régulière pour $f^T L f$, les vecteurs propres associés aux plus petites valeurs propres sont les plus intéressants. Il existe des algorithmes implémentés (package RSopectra de R, ...) pour ce problème acceptant de grandes matrices en entrée.

Ici, nous proposons plutôt une inversion basée sur la méthode de Choleski, après changement de base pour se ramener au problème de l'inversion exacte d'une sous-matrice symétrique définie positive. L'algorithme d'orthonormalisation de Gram-Schmidt est explicité dans la proposition suivante. L'objectif, partant du vecteur $\mathbf{1}$ (qui est un vecteur propre de L associé à la valeur propre zéro) est de le compléter par $N-1$ vecteurs libres formant une base (par exemple e_j est tel que $e_j(i) = 0$ sauf $e_i(j-1) = 1$ et $e_j(j) = -1$), puis d'en déduire une base orthonormée associée à une matrice orthogonale de changement de base, dans laquelle on pourra exprimer L en blocs.

Proposition 4.1.1. *Soit L le Laplacien du graphe G, connecté, non dirigé. L ayant une seule valeur propre nulle, associé au vecteur propre $\mathbf{1}$. La méthode itérative d'orthonormalisation de Gram-Schmidt (ref) permet d'écrire que*

$$P = \begin{pmatrix} \frac{1}{\sqrt{N}} & 1/\sqrt{2} & \dots & 1/\sqrt{k(k-1)} & \dots & \dots & 1/\sqrt{N(N-1)} \\ \frac{1}{\sqrt{N}} & -1/\sqrt{1-1/2} & \vdots & 1/\sqrt{k(k-1)} & \vdots & \vdots & 1/\sqrt{N(N-1)} \\ \frac{1}{\sqrt{N}} & 0 & \ddots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{N}} & 0 & 0 & -1/\sqrt{1-1/k} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 1/\sqrt{N(N-1)} \\ \frac{1}{\sqrt{N}} & 0 & 0 & 0 & \dots & 0 & -\sqrt{1-1/N} \end{pmatrix}$$

est une matrice orthogonale de changement de base dont le premier vecteur est constant. Donc la matrice $L' = P^T L P$ exprimant l'application linéaire L dans la nouvelle base admet la représentation en blocs diagonaux

$$L' = P^T L P = \left(\begin{array}{c|c} 0 & 0 \cdots 0 \\ \hline \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & A \end{array}\right)$$

où A est une matrice symétrique inversible de taille $N_v - 1$. Après inversion de A, on peut écrire la pseudo-inverse de L comme :

$$L^+ = P \left(\begin{array}{c|c} 0 & 0 \cdots 0 \\ \hline \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & A^{-1} \end{array}\right) p^T$$

Même pour de grandes matrices (de l'ordre de 60 000 lignes), il n'est pas compliqué de construire la matrice L^+ si assez de mémoire vive est disponible. Cela nécessite 2 produits de matrices carrées. Ensuite, l'inversion de A a la même complexité que ces produits. Il s'est avéré que cette méthode permettait d'exhiber la matrice pseudo inverse L^+ sans avoir à diagonaliser L, ce qui n'est pas accessible pour des matrices de cette taille. Les éléments numériques sont décrits dans la section suivante.

4.1.6 Réflexion sur la complexité algorithmique des méthodes

Pour des matrices pleines (sans codage en matrices creuses et leurs algorithmes adaptés), on cite ci-dessous quelques informations utiles pour le passage à l'échelle de grandes matrices. Le choix (discutable) de ne pas adapter les méthodes aux matrices creuses vient d'une contrainte matérielle : les librairies d'algèbre linéaire BLAS et LAPACK dans R sont utilisées par le package "Matrix" de Douglas Bates et Martin Maechler ([Bates, 2009], [Douglas Bates, 2019]), qui étend les fonctions aux matrices creuses de tous types, mais ces fonctions imposent que les indices des matrices soient codées sur 32 bits. C'est à dire que dans sa version actuelle (2019), une matrice ne peut pas avoir plus de $2^{31} - 1$

éléments, soit pour une matrice carrée une taille maximale de 46340 lignes. En outre, même si le Laplacien est une matrice très creuse, son inverse sera une matrice dense.

Plus généralement, sans faire une analyse exhaustive et détaillée des problèmes numériques associés aux calculs sur de grandes matrices carrées, nous donnons quelques éléments de compréhension pour les ordres de grandeur, à la fois en capacité de stockage mais aussi en nombre d'opérations (et donc de temps de calcul). Ben sûr, quand c'est possible (déjà prévu par les fonctions implémentées), la parallélisation sur plusieurs cœurs permet de gagner beaucoup de temps. Nous donnons ainsi les ordres de grandeur des temps de calculs tels qu'ils ont été effectués sur le cluster PLAFRIM, supporté par Inria, le CNRS (LABRI and IMB), l'université de Bordeaux, Bordeaux INP et le Conseil Régional d'Aquitaine (voir https ://www.plafrim.fr/). Les calculs ont été lancés sur les nœuds *miriel* de PLAFRIM2, à 128 Go de RAM et 24 cœurs, processeurs INTEL Xeon (2,5 GHz) [¶].

1. Produit de deux matrices carrées de taille n : nombre théorique d'opérations de l'ordre de $0(n^3)$ à $O(n^{2,807})$ (algorithme de Strassen), prend environ 15 minutes par produit pour des matrices carrées de taille 60 000. Chaque matrice requiert une capacité de stockage de 24,5Go.
2. Factorisation de Choleski : pour une matrice réelle A, symétrique, définie positive (donc inversible), l'objectif est de déterminer l'unique matrice triangulaire inférieure B telle que $A = BB^T$ et $diag(B) > 0$. La complexité est en $n^3/3 + O(n^2)$.
3. Inversion d'une matrice réelle symétrique par la méthode de Choleski : nombre théorique d'opérations de l'ordre de $O(n^3)$ à $O(n^{2,807})$. Cela prend environ 7 heures pour une matrice de taille 60 000. Le calcul explicite de l'inverse A^{-1} a un cout similaire à la factorisation de Choleski, mais une durée en générale (juste) trois fois plus grande : $O(n^3)$ pour la décomposition BB^T, puis n étapes de montées et descentes de systèmes triangulaires, chacun en $O(n^3)$.
4. Diagonalisation de matrices symétriques : déterminer le polynôme caractéristique et calculer ses racines est une méthode numériquement instable et donc inadaptée pour obtenir les valeurs propres d'une matrice. L'algorithme le plus fiable semble être la méthode QR de recherche de valeurs propres, basée sur la factorisation de Householder. C'est une méthode itérative, dont chaque itération a un cout $O(n^3)$. Il est donc difficile de donner une évaluation de son cout mais cela justifie l'impossibilité d'obtenir la totalité du spectre d'une grande matrice pleine de grande taille.
5. Recherche d'une partie des valeurs propres : la méthode complète est coûteuse en temps mais aussi en mémoire car elle demande de stocker des matrices triangulaires et de Hessenberg de la même taille que A. De même en termes de nombre d'opérations, elle a une complexité en $O(n^3)$, mais une constante bien supérieure aux méthodes de multiplication et d'inversion, ce qui peut s'avérer rédhibitoire pour des matrices de la taille de quelques dizaines de milliers de lignes. Pour cette raison, des méthodes alternatives sont utilisées lorsque seul un sous-ensemble du spectre est désiré. C'est le cas des méthodes implémentées dans le package RSpectra ([Qiu and Mei,], basées sur des méthodes de puissance. Ce package, qui a été

¶. documentation matérielle de PLAFRIM : https ://www.plafrim.fr/hardware-documentation/

testé sur de grandes matrices laplaciennes fonctionne bien dans R, il est basé sur des algorithmes de Lehoucq et Sorenson ([Lehoucq et al., 1998]) [‖].

On retiendra donc qu'inverser une matrice carrée de taille 60 000 prend environ 7 heures (avec une utilisation maximale des 24 cœurs). Quand la méthode ne requiert qu'une seule inversion, cet effort peut être envisagé. Chaque produit de deux matrices carrées prend 15 à 30 minutes. Nous avons utilisé le langage R (R3.6.2) pour nos calculs. R a l'avantage de contenir de très nombreuses librairies statistiques, mais aussi de calculs numériques. R peut aussi manipuler les matrices creuses, améliorant potentiellement grandement les temps de calculs une fois réglés les problèmes du package Matrix.

4.2 Proposition pour un choix automatique de λ

Comme toute méthode régularisée, la régression ridge à noyau utilisée ici nécessite de choisir un coefficient de pondération entre l'attache aux données et la régularité de la fonction. On rappelle brièvement les deux méthodes de validation croisée, et de validation croisée généralisée, cette dernière méthode ayant été utilisée sur nos cas d'application.

Il existe plusieurs stratégies possibles en validation croisée pour évaluer la performance d'un modèle afin de définir les paramètres de pénalisation optimaux (en un sens à définir). Dans l'ensemble elles consistent à diviser les données en deux groupes. Le premier appelé 'entrainement' qui joue le rôle de données initiales sur lesquels le modèle est construit. Le second 'test' qui joue le rôle de nouvelles données et qui est utilisé pour évaluer la performance du modèle construit sur le jeu de test. Ainsi l'évaluation de la performance d'un modèle obtenu dépend de la façon dont le jeu de donnée initial est divisé.

La méthode *k-fold-cross-validation* consiste à diviser le jeu de données en k échantillons de tailles plus ou moins équivalentes et mutuellement exclusifs et permet d'éviter la dépendance entre l'évaluation de la performance du modèle et la division du jeu de données. La validation croisée "leave-one-out" (LOOCV) est un cas particulier où $k = n$, c'est-à-dire que l'on apprend sur $n-1$ observations puis on valide le modèle sur la n-ième observation et l'on répète cette opération n fois. On décrit ci-dessous la LOOCV ainsi que la validation croisée généralisée qui a finalement été employée.

4.2.0.a La validation croisée LOOCV

Dans la procédure LOOCV, on définit le prédicteur $\hat{f}_{-i}(\lambda) \in \mathbb{R}^{N_v}$ appris sur l'échantillon $V_{obs,(-i)} = \{j \in V_{obs}, j \neq i\}$ et la valeur spécifique $\hat{f}_{i(-i)}$ obtenu sur le nœud $v(i)$. Ainsi, le i-ème résidu *leave-one-out* $z_i - \hat{f}_{i(-i)}$ permet de construire la statistique PRESS (Predicted Error sum of squares), qui est minimisée en λ pour définir le critère de sélection par validation croisée :

$$\lambda_{opt,LOOCV} = \arg\min_{\lambda} \frac{1}{n} \sum_{i=1}^{n} [z_i - \hat{f}_{i-(i)}(\lambda)]^2$$

On rappelle que le noyau $\boldsymbol{K}$ admet la représentation (voir 4.1.22)

‖. (voir aussi la thèse de Lehoucq : [Lehoucq, 1995])

$$K = \Psi^T \Psi = \sum_{i=1}^{N_v} \psi_i^T \psi_i$$

où la matrice Ψ admet en colonne les vecteurs

$$\psi_i = \psi(v_i) = \begin{pmatrix} \sqrt{\delta_1}\phi_1(i) \\ \sqrt{\delta_2}\phi_2(i) \\ . \\ . \\ \sqrt{\delta_{N_v}}\phi_{N_v}(i) \end{pmatrix} \tag{4.2.1}$$

Si on veut obtenir le prédicteur *leave-one-out* de f_i, il faut ajuster la régression ridge sur l'échantillon $V_{obs,(-i)} = \{j \in V_{obs}, j \neq i\}$ donc résoudre le problème

$$min_{f \in \mathcal{H}} \sum_{j \in V_{obs}, j \neq i} (x_j - f(j))^2 + \lambda \|f\|_{\mathcal{H}}^2$$

qui, d'après le théorème 4.1.3 admet une représentation de la forme :

$$\forall v \in V, f(v) = \sum_{j=1, j \neq i}^{n} \alpha_j K(j, v),$$

codée vectoriellement

$$f_{(i)} = \boldsymbol{K}_{(-i)}^{N_v,(n-1)} \alpha_{(-i)}$$

où $\alpha_{(-i)} \in \mathbb{R}^{n-1}$ et $\boldsymbol{K}^{N_v,(n-1)}$ est la matrice de taille $N_v \times (n-1)$ qui contient les $(n-1)$ colonnes de $\boldsymbol{K}$ associées aux nœuds observés $V_{obs,(-i)}$.

L'ajustement sur l'échantillon $V_{obs,(-i)}$ donne donc le vecteur restreint à $V_{obs,(-i)}$ des valeurs prédites

$$\hat{f}_{obs,(-i)} = \boldsymbol{K}_{-i,-i}^{(n-1)} \hat{\alpha}_{(-i)} \in R^{(n-1)}$$

qu'on étend à tout le graphe :

$$\hat{f}_{(-i)} = \boldsymbol{K}_{(-i)}^{N_v,(n-1)} \hat{\alpha}_{(-i)} \in \mathbb{R}^{N_v},$$

et dont la norme vaut

$$\|\hat{f}_{(-i)}\|_{HH} = \hat{\alpha}_{(-i)}^T \boldsymbol{K}_{-i,-i}^{(n-1)} \hat{\alpha}_{(-i)}^T$$

où $\boldsymbol{K}_{-i,-i}^{(n-1)}$ est le noyau évalué sur $V_{obs,(-i)}$.

L'application de la régression ridge à noyau donne donc

$$\hat{\alpha}_{(-i)} = \left(\boldsymbol{K}_{-i,-i}^{(n-1)} + \lambda I_{(n-1)}\right)^{-1} z_{obs,(-i)}, \tag{4.2.2}$$

et la i-ème prédiction *leave-one-out* est

$$\hat{f}_{i,(-i)} = \boldsymbol{K}_{i,(-i)}^{N_v,(n-1)} \left(\boldsymbol{K}_{-i,-i}^{(n-1)} + \lambda I_{(n-1)}\right)^{-1} z_{obs,(-i)}$$

où $\boldsymbol{K}_{i,(-i)}^{N_v,(n-1)}$ est la ligne i du noyau $\boldsymbol{K}_{(-i)}^{N_v,n-1}$ et $I_{(n-1)}$ est la matrice identité de taille $(n-1)$.

Remarque sur la LOOCV rapide et la formule de Woodbury

La méthode LOOCV requiert théoriquement la répétition de n régressions Ridge à noyau sur des échantillons de taille $n-1$. Le cout computationnel peut être prohibitif pour de grands jeux de données, surtout quand chaque évaluation est elle-même couteuse. C'est habituel pour les méthodes de validation croisée *leave-one-out* dont le paramètre estimé est de la forme

$$\hat{\theta} = \left(X^T X + \lambda I\right)^{-1} X^T Y,$$

pour un modèle de type $Y = X\theta + \epsilon$, l'estimateur des moindres carrés étant obtenu pour $\lambda = 0$. Quand les **colonnes** de la matrice de design X sont fixées (ce qui est le cas en régression linéaire standard), le calcul du critière CV ne nécessite finalement qu'une seule évaluation sur le jeu de données entier car l'estimateur *leave-one-out* $\hat{\theta}_{(-i)}$ de θ obtenu en enlevant la ligne i de X peut s'écrire à partir de l'estimateur $\hat{\theta}$ grace à la fameuse formule de Sherman–Morrison (une version simplifiée de l'égalité de Woodbury, [Hager, 1989]) permettant l'astuce matricielle[**]

$$(A + uv^T)^{-1} = A^{-1} + \frac{A^{-1}uv^T A - 1}{1 + v^T A^{-1} u}.$$

Ainsi, comme $X_{(-i)}^T X_{(-i)} = X^T X - x_i x_i^T$, chaque évaluation de $\hat{\theta}_{(i)}$ peut être obtenue sans avoir à inverser une nouvelle matrice mais simplement en ajustant le résultat global par des produits matriciels peu coûteux, voir [van Wieringen, 2015a] pour plus de détails.

Malheureusement, la régression ridge à noyau ne semble pas entrer dans ce cadre, car l'estimation *leave-one-out* sur le nœud i non seulement supprime la ligne i du noyau $\boldsymbol{K}^{(n)}$, mais également la i-ème colonne, d'après le théorème du représentant, comme illustré équation (4.2.2). Pourtant ([Elisseeff et al., 2003], Lemme 4.1) utilise cet argument sans justification pour proposer une formule simple du résidu *leave-one-out* sans recalcul :

$$z_i - f_{i(-i)} = \frac{z_i - f_i}{1 - H_{ii}}$$

où $H = \boldsymbol{K}^{(n)}(\boldsymbol{K}^{(n)} + \lambda I)^{-1}$ est la Hat matrice définie en (4.1.20). Quelques calculs numériques de vérification sur un graphe ont montré que cette formule est fausse et ne s'applique pas dans notre cas. La validation croisée LOOCV n'est donc pas une méthode applicable en pratique pour la régression ridge à noyau sur de grands graphes.

4.2.0.b Validation croisée généralisée

La validation croisée généralisée est une autre méthode permettant de guider le choix du paramètre de pénalité λ. On procède comme en validation croisée mais avec un critère différent pour évaluer les performances de l'estimateur ridge sur les nouvelles données. Ce critère, noté GCV(λ) [Golub et al., 1979], (pour Generalised Cross-Validation, en anglais) est une approximation de la statistique PRESS de Allen [Allen, 1974].

On a vu que dans le cadre standard de la régression ridge, mais pas pour notre régression à noyau, il est possible d'éviter de refaire une nouvelle régression à chaque i pour

**. que l'on applique à $A = X^T X + \lambda I$, et $u = v = x_i$ la i-ème ligne de X.

prédire la valeur en i, sans le point i. Il est possible de reformuler l'estimateur comme présenté dans [van Wieringen, 2015b] en utilisant la matrice identité de Woodbury et un ré-écriture du prédicteur. La statistique de Allen PRESS peut alors s'écrire de la forme :

$$\lambda_{opt} = \arg\min_{\lambda} \frac{1}{n} \sum_{i=1}^{n} [z_i - \hat{f}_{i(-i)}(\lambda)]^2 = \arg\min_{\lambda} \frac{1}{n} \parallel B(\lambda)[\mathbf{I} - H(\lambda)]z_{obs} \parallel^2 \tag{4.2.3}$$

où $B(\lambda)$ est une matrice diagonale avec $B(\lambda)_{i,i} = [1 - H_{i,i}(\lambda)]^{-1}$. Par conséquent, les performances de prédiction pour un λ donné peuvent être évaluées depuis la *hat matrix* $\hat{H}$ et le vecteur réponse Y sans recalculer les k *leave-one-out* estimateurs ridges.

Cette formule générale de validation croisée est à la base d'une adaptation qui propose de remplacer la matrice diagonale $B(\lambda)$ par sa "moyenne" et éviter la dispersion des poids diagonaux $H_{ii}(\lambda)$ dont certains peuvent prendre des valeurs proches de 1. Cela donne en effet un trop gros poids aux noeuds i pour qui $1/(1(-H_{ii}(\lambda))$ pourraient prendre de très grandes valeurs. Comme $tr[H(\lambda)] = \sum_{i=1}^{n} H_{ii}(\lambda)$, une méthode pour stabiliser les poids à une valeur constante est de remplacer chaque $H(\lambda)_{ii}$ par $\frac{1}{n} tr(H(\lambda))$.

Le critère GCV ([Golub et al., 1979]) est alors défini par

$$GCV(\lambda) = \frac{\frac{1}{n} \parallel (\mathbf{I} - H(\lambda))z_{obs} \parallel^2}{[\frac{1}{n} tr(\mathbf{I} - H(\lambda))]^2}$$

L'estimateur GCV de λ est alors

$$\lambda_{opt,GCV} = \arg\min_{\lambda \in \mathbf{R}^+} GCV(\lambda)$$

Dans notre cas la *hat matrix* est la matrice d'influence $H(\lambda) = \boldsymbol{K}^{N_v,n}(\boldsymbol{K}^{(n)} + \lambda I)^{-1}$.

Pour conclure, la recherche d'un paramètre de lissage λ passe forcément ici par un calcul d'inversion du Laplacien de graphe puisque la méthode de validation croisée généralisée requiert la connaissance du noyau $\boldsymbol{K}$, et l'inversion (ou la diagonalisation) d'une extraction $\boldsymbol{K}^{(n)}$. Les méthodes de type calcul direct du prédicteur par descente de gradient pour $\hat{f}$ ne permettent pas cette analyse. Cela donne un avantage à l'effort exceptionnel d'inversion du Laplacien.

4.3 Développement : le signal à reconstruire est une série temporelle

Dans un objectif de généralisation, et parce que l'application industrielle est adaptée à ce cas, on suppose maintenant que des observations sont disponibles sous la forme de séries temporelles pour un sous-ensemble de nœuds $V^{obs} \subseteq V$. Plus précisément, pour tout $i \in V^{obs}$, on dispose d'un vecteur $\boldsymbol{z}_i \in \mathbb{R}^M$ représentant les données observées aux temps $t_1 < t_2 < \ldots < t_M$.

Notre objectif principal est de reconstruire les valeurs inconnues sur l'ensemble du graphe $z_i, i \notin V^{obs}$ pour lesquelles nous n'avons pas d'observation. On notera comme d'habitude par $n = |V^{obs}|$ le nombre de nœuds contenant des observations. Le graphe $G = (V, E)$ et son Laplacien $L = D - A$ ne sont pas supposés varier au cours du temps.

4.3.1 Espace de Hilbert à noyau reproduisant de fonctions à valeurs vectorielles

Dans cette section, nous rappelons des notions de base sur l'apprentissage statistique dans l'espace de Hilbert à noyau reproduisant des fonctions à valeurs vectorielles. En se basant sur les travaux de [Micchelli and Pontil, 2005b], [Micchelli and Pontil, 2005a], [Alvarez et al., 2011] et [Romero et al., 2017], où les espaces de Hilbert de fonctions à valeurs vectorielles sont introduits à des fins d'apprentissage multitâche, nous proposons une approximation d'un signal temporel discret sur le graphe à partir d'un échantillon de signaux observés sur la même période de temps. Il ne s'agit pas de prédiction future mais de reconstruction en tout point du réseau d'un vecteur de valeurs mesurées en $t_1 < t_2 < \cdots < t_M$.

Pour simplifier les notations, on supposera dans la suite que la structure du graphe n'évolue pas avec le temps. La méthode naïve de reconstruction *temps par temps* est alors la simple répétition M-fois des méthodes de régression ridge vues précédemment. Afin de proposer une méthode globale qui tienne compte d'une éventuelle régularité temporelle du signal, on présente ci-dessous une technique de reconstruction à noyau spatio-temporel, basés sur les RKHS à valeurs vectorielles. Malheureusement, l'exploitation complète de ces techniques demande des capacités numériques beaucoup plus grosses que celles déjà employées. Il a donc fallu réduire les ambitions pour l'application à de grands réseaux.

4.3.1.a Noyaux pour un RKHS à valeurs vectorielles

Soit $\tilde{\mathcal{H}}$ un espace de fonctions réelles $f : \mathcal{V} \times \mathbb{R}^M \to \mathbb{R}$, codant en $f(v_i, t_k)$ la valeur de f observée sur le noeud $v_i \in \mathcal{V}$ au temps t_k. Notons que c'est équivalent à définir $\mathcal{H}$ l'espace des fonctions $\boldsymbol{f}$ à valeurs vectorielles $\mathcal{V} \to \mathbb{R}^M$ codant dans le vecteur $\boldsymbol{f}(v_i) = (f(v_i, t_1), \ldots, f(v_i, t_M))^T \in \mathbb{R}^M$ les M observations au nœud v_i.

Dans le même esprit qu'Aronzaj (1950), à tout espace de Hibert de fonctions à valeurs vectorielles correspond un noyau, que nous allons définir. Également, à chaque noyau correspondra un RKHS unique (à une isométrie près).

En tant que généralisation du cas univarié $M = 1$, le noyau associé à la structure du graphe qui va permettre de capter la régularité supposée du signal dans le temps devra lisser à la fois spatialement la fonction en tout temps t_k sur les noeuds connectés par une arrête, mais aussi pénaliser les grands écarts de f sur chaque noeud en des temps voisins t_{k-1}, t_k, t_{k+1}. Autant le choix du Laplacien du graphe était un candidat naturel pour construire le noyau K (avec éventuellement ses versions plus régularisées $K = \Phi\, r(\Delta)\, \Phi^T$), autant le choix du lissage temporel offre une grande liberté, contrainte seulement par des critères de calculabilité en temps raisonnable pour des graphes de grandes tailles.

Soit $\mathcal{M}_M(\mathbb{R})$ l'ensemble des matrices réelles de taille M, et $\mathcal{M}_M^+(\mathbb{R})$ le sous ensemble des matrices semi-définies positives.

Définition 4.3.1. K est un noyau de $\mathcal{V} \times \mathcal{V} \to \mathcal{M}_M(\mathbb{R})$ s'il satisfait, pour tout $(i, k) \in V \times \mathcal{V}$, les propriétés suivantes :

(a) $K(i, j) = K(j, i)^T$ et $K(i, i) \in \mathcal{M}_M^+(\mathbb{R})$.

(b) Pour tout $m \in \mathbb{N}$, $\{i_1, \ldots, i_m\} \subset \mathcal{V}$, $\{z_j : 1 \leq j \leq m\} \subset \mathbb{R}^M$ on a

$$\sum_{\ell,\ell'=1}^{m} \langle z_\ell, K(i_\ell, i_{\ell'}) z_{\ell'} \rangle_{\mathbb{R}^M} \geq 0.$$

On définit l'espace de Hilbert $\mathcal{H}$ à valeurs vectorielles associé au noyau reproduisant K. Ici en dimension finie, toute fonction de cet espace s'écrit comme combinaison linéaire de fonctions $K(x_i, .)$

$$f(x) = \sum_{i=1} K(x_i, x) c_i, \ c_i \in \mathbb{R}^M.$$

De plus, $\mathcal{H}$ a la propriété d'être un espace à noyau reproduisant, c'est à dire que, $\forall c \in \mathbb{R}^M, x \in \mathcal{V}$, $K(x, x')c$ comme fonction de x' appartient à $\mathcal{H}$ et

$$< f, K(x, .)c >_{\mathcal{H}} = f(x)^T c.$$

L'intérêt de placer le problème dans le cadre des RKHS à valeurs vectorielles est qu'on obtient directement une expression du minimiseur de l'erreur quadratique régularisée : on cherche ici à minimiser dans $\mathcal{H}$

$$\hat{\boldsymbol{f}} = \arg\min_{f \in \mathcal{H}} \sum_{i \in V^{obs}} \|\boldsymbol{z}_i - \boldsymbol{f}(i)\|^2_{\mathbb{R}^M} + \lambda \|\boldsymbol{f}\|^2_{\mathcal{H}}. \tag{4.3.1}$$

A partir du théorème 4.1 dans [Micchelli and Pontil, 2005a], le problème d'optimisation ci-dessus admet une unique solution et son calcul revient à résoudre un système d'équations linéaires à dimensions finies.

Proposition 4.3.1. *L'estimateur $\hat{\boldsymbol{f}}$ admet la forme suivante*

$$\hat{\boldsymbol{f}} = \sum_{i \in V^{obs}} K(x_i, .) c_i,$$

où les coefficients $c_1, \ldots, c_n$ sont des vecteurs dans $\mathbb{R}^M$ qui sont la solution unique du système linéaire d'équations

$$(G + \lambda I)c = \boldsymbol{z}, \tag{4.3.2}$$

où c est la concaténation des $(c_i)_{i \in V^{obs}}$ et $\boldsymbol{z}$ est la concaténation des $(\boldsymbol{z}_i)_{i \in V^{obs}}$, tous deux vecteurs dans $\mathbb{R}^{nM}$, I est la matrice d'identité de taille $nM \times nM$, et G une matrice par bloc, chaque bloc étant de taille $n \times n$ tel que son j-ème, k-ème bloc est $G_{jk} = K(j, k) \in \mathcal{M}_M(\mathbb{R})$ (ainsi G est une matrice réelle de taille $nM \times nM$).

Bien sûr, ces notations sont cohérentes avec le cas $M = 1$, auquel cas, chaque "bloc" est de taille 1 et on retrouve le noyau K initial tiré du Laplacien de graphe.

4.3.1.b Choix de noyau pour un graphe statique et des mesures temporelles

Le cas d'étude concerne un graphe n'évoluant pas dans le temps, et un suivi complet d'un échantillon de nœuds $i \in V_{obs}$ sur un ensemble M de mesures aux temps $t_1 < \cdots < t_M$. Afin de respecter les connexions entre nœuds en tout temps, il est naturel de supposer

dans un premier temps une analyse *temps par temps* pour laquelle chaque estimation du signal au temps t_k tient compte uniquement des informations au temps t_k, c'est à dire en répétant M fois une estimation à valeurs réelles comme dans les parties précédentes. Ce n'est évidemment pas satisfaisant pour régulariser le signal dans son aspect temporel, mais on peut imaginer des résultats cohérents si le signal lui-même est déjà assez régulier (avec une fréquence d'observations suffisamment élevée au regard de la variabilité temporelle).

Afin de tenter de tenir compte des liaisons éventuelles entre les valeurs du signal en des temps différents, [Romero et al., 2017] partent de la définition d'un graphe *étendu*, c'est à dire un graphe "spatio-temporel" $\tilde{G} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ en répliquant M fois les nœuds $v \in V_{obs}$, c'est à dire en définissant le nœud $v[t]$ pour $v \in V$ et $t \in \{t_1, \dots t_M\}$. Les arrêtes sont aussi à définir et basiquement, tous les nœuds d'un même temps t partagent la structure du graphe initial (avec sa matrice d'adjacence valuée A), supposée constante dans le temps. Par contre, les connexions entre nœuds à des temps différents $t_k, t_{k'}$ sont à définir.

Cette super matrice d'adjacence $\tilde{A}$ de taille $N_v M$ peut prendre des formes assez simples, par exemple (par degré de complexité croissant) :

1. une matrice $\tilde{A}$ diagonale par bloc, chaque bloc de taille N_v étant égal à la matrice d'adjacence A du graphe $\mathcal{G}$.
2. une matrice $\tilde{A}$ tridigonale par bloc, chaque bloc diagonal de taille N_v étant égal à la matrice A, puis les sous- et sur-diagonales recevant une matrice carrée B diagonale de taille N_v stockant la connexion à un temps de décalage entre les nœuds v_i et v_j, avec $b_{ij} = b\delta_{ij}$. Son interprétation repose simplement sur l'hypothèse qu'entre deux temps successifs, chaque noeud v est connecté avec lui-même par un poids de force b. Il existe une arrête entre $v_i[t]$ et $v_i[t-1], i \in calV$. Il est alors crucial de considérer la valeur de b relativement aux poids w_{ij} de la matrice d'adjacence initiale A.
3. une matrice $\tilde{A}$ tridigonale par bloc, chaque bloc diagonal de taille N_v étant égal à la matrice A, puis les sous- et sur-diagonales recevant une matrice carrée B quelconque de taille N_v stockant la connexion à un temps de décalage entre les nœuds v_i et v_j.

Cette approche consiste finalement à agglomérer les deux dimensions (spatiale et temporelle) pour définir directement une matrice d'adjacence, à l'origine de la définition d'un Laplacien, puis d'un noyau.

La première forme décrite revient à analyser le signal "temps par temps", sans valoriser l'aspect lisse de la fonction dans le temps.

La seconde forme (ainsi que la suivante) nécessite de calculer un nouveau noyau à partir d'un Laplacien de très grande taille ($N_v M$). Malheureusement, pour un signal "haute fréquence", il n'est pas raisonnable avec les moyens à disposition d'essayer de résoudre le problème d'inversion d'une matrice de taille de l'ordre du milliard de lignes : le noyau du réseau de Bordeaux fait 57 400 nœuds, la résolution pour un signal mesuré toutes les 5mn sur 2 jours demande l'inversion d'une matrice de taille 57 400*12*24*2=33 062 400 !

On peut par contre imaginer développer une méthode par descente de gradient, si le stockage simple de cette matrice est accessible. Pour des routines d'algèbre linéaire stockant la matrice entièrement en mémoire vide, il faut réserver au moins 7,5 To de RAM. Des développements supplémentaires sont donc attendus pour simplifier le problème.

4.4 Applications numériques

4.4.1 Quel signal reconstruire ?

Dans cette section nous présentons la reconstruction de différents signaux définis sur les nœuds d'un graphe à partir de la méthode d'inférence présentée dans ce chapitre. Nous avons à notre disposition trois signaux différents à reconstruire :

1. **Un signal simulé lisse à basse fréquence (bandlimited) :** le but de ce signal est d'évaluer les capacités de reconstruction de la méthode d'inférence sur le cas concret des graphes représentant les réseaux AEP *Bx* et *Ambès*. Les données sur les nœuds du graphe de ce signal sont entièrement simulées Section 4.4.2.
2. **Le signal reconstruit du débit :** les données de ce signal sont reconstruites grâce sur tous les nœuds du graphe à l'algorithme d'optimisation du Chapitre 3 à partir de données réelles. Ce signal nous permet de tester la reconstruction ayant un sens d'un point de vue équilibre hydraulique du réseau Section 4.4.3.
3. **Le signal de chlore :** représente les mesures de concentration de chlore (en $^{mg}/_{L}$) effectué par un très faible nombre de capteurs qualités sur le réseau. Les données considérées ici sont réellement observées sur les nœuds du graphe Section 4.4.4.

Les deux premiers signaux étant définis (simulés ou reconstruits) sur tous les nœuds du graphe nous sommes en mesure d'évaluer la qualité de reconstruction des méthodes d'inférence à noyau sur ces réseaux. Le signal de chlore quant à lui n'est mesuré que partiellement sur le graphe, la reconstruction du signal ne peut s'effectuer qu'a partir d'un nombre restreint d'observations.

Ces signaux sont définis sur les deux réseaux *Ambès* et *Bx*. Celui d'*Ambès* de petite taille nous permet de tester l'ensemble des méthodes avec un coût computationnel faible. Celui de *Bx* de très grande taille nous oblige à utiliser de plus gros moyens de calcul dans le but de valider les méthodes.

Afin de tester les méthodes, plusieurs niveaux d'échantillonnages aléatoire des nœuds considérés comme observés ont été testés.

Dans le cas réelles nœuds observés V_{obs} représentent les positions des capteurs qualités équipés sur les réseaux AEP et représentent moins de 1% du total de nœuds sur le réseau de graphe $p = \frac{|V_{obs}|}{N_v} \times 100 = 0.88\%$ ce qui est très peu pour les méthodes de lissages utilisées. Dans le cas simulé les nœuds sont choisis de façon aléatoire par tirage uniforme sans remise de l'ensemble des nœuds. Différents niveaux d'échantillonnage sont testés $p = \{10\%, 20\%, 40\%, 60\%\}$. Pour chacun de ces tirages les nœuds étant réellement équipés de capteurs sont ajoutés à la liste des noeuds échantillonnés. La Figure 4.1 représente la position des nœuds étant considéré comme des observations (en rouge) lors de l'estimation des débits. La projection des noeuds observés pour le graphe *Bx* est présentée en Annexe G.

La *Hat matrix* globale H (Equation 4.1.20), nous renseigne sur la façon dont chaque valeurs de $f_i, i \in V$ est reconstruite comme une combinaison particulière des observations $x_{obs}(i)$. Ainsi la i^{eme} ligne de H contient les coefficients à appliquer à x_{obs} pour calculer $\hat{f}_i$. Les deux graphes Figure 4.2 montrent les coefficients H_{600} et H_{1500} appliqués à chacune des observations pour reconstruire le signal sur les noeuds $i = 600$ et $i = 1500$ représentés en violet sur le graphe de gauche et en rose sur le graphe de droite. Plus le nœud observé est proche du nœud à estimer, plus la valeur du coefficient est élevée.

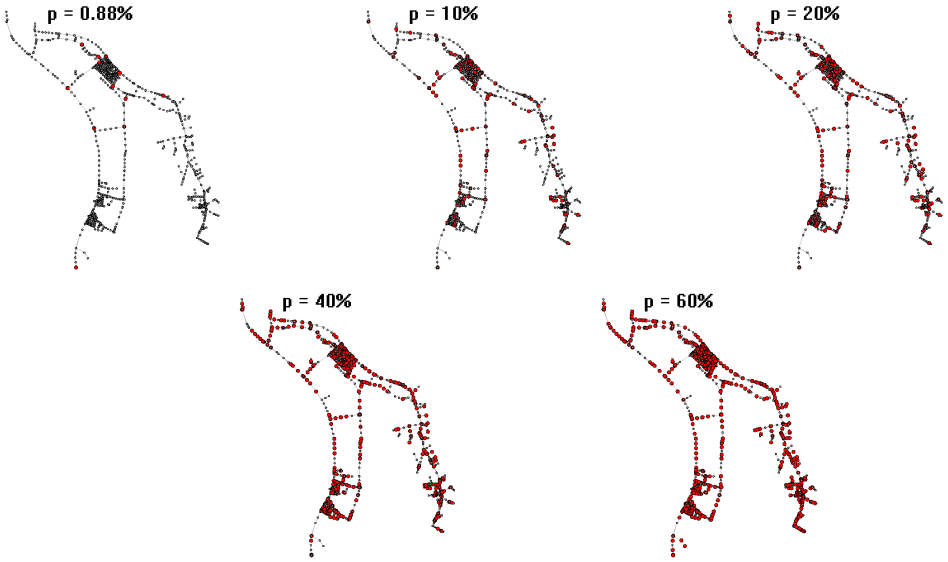


FIGURE 4.1 – Projection sur les noeuds du graphe *Ambès* des nœuds observés V_{obs} en rouge pour différents niveaux d'échantillonage p

A l'inverse les colonnes de la *Hat matrix* H nous informent sur l'influence des points de mesures sur la reconstruction du signal. La Figure 4.3 fournit une représentation visuelle des coefficients en colonne $H_{.,j}$ dans le cas réel où $p = 0.88\%$ des nœuds considérés comme observés. La valeur des coefficients $h_{i,j}$ de la *hat matrix* est représentée par un gradient de couleurs. On observe ainsi l'impact de la valeur observée sur l'ensemble des nœuds du graphes, avec un coefficient $h_{i,j}$ élevé pour les nœuds proches du nœuds observés et plus faible à mesure que l'on s'éloigne de celui-ci.

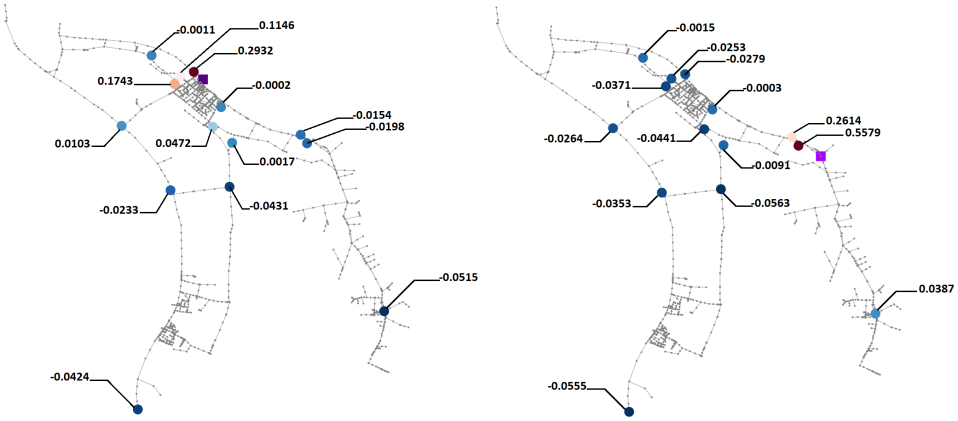


FIGURE 4.2 – Coefficients $h_{(i,j)} \forall j \in V_{obs}, i \in \{600, 1500\}$ à appliquer pour la reconstruction individuelle du noeud $i = 600$, resp. $i = 1500$, en violet, resp. rose sur les graphes, dans le cas $p = 0.88\%$.

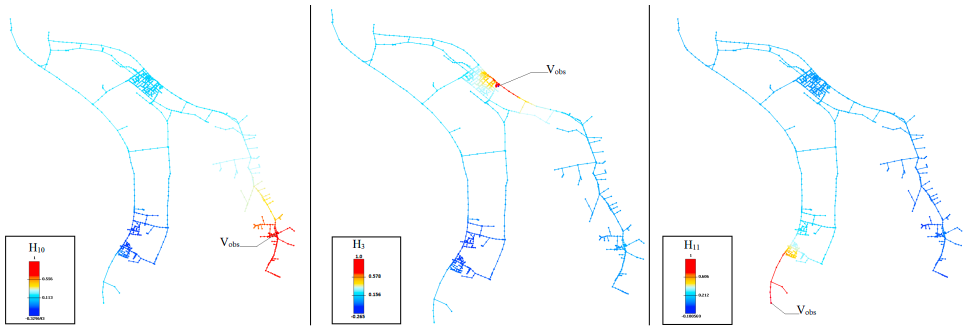


FIGURE 4.3 – Influence des poids de mesures $h_{(i,j)} \forall i \in \{1, \ldots, N_v\}$, $j = 10, 3, 11$ de gauche à droite sur la reconstruction, dans le cas $p = 0.88\%$. Le nœud considéré comme observé est pointé sur chaque graphe par le symbol V_{obs}. Pour plus de lisibilité la couleur des nœuds est interpolée sur les arêtes.

4.4.2 Un signal simulé band-limited

Le signal que l'on cherche à reconstruire ici est un signal obtenu comme une combinaison des k premiers vecteurs propres du noyau K, c'est à dire les derniers non nuls de la matrice Laplacienne $\boldsymbol{L}$ du graphe. L'idée est d'obtenir un signal band-limited comme une combinaison linéaire des premiers vecteurs propres associés au noyau $K = \boldsymbol{L}^{-1}$ de la forme :

$$Y = \sum_{i=2}^{k} \phi_i \times c_i$$

avec $\phi_i, \forall 2 \leq i \leq k$, les vecteurs propres de $\boldsymbol{L}$ associés aux plus petites valeurs propres de K et $c_i \in \mathbb{R}^+$ des coefficients donnés avec $c_i > c_{i+1}, \forall i \leq k$ et $c_i = 0$ si $i > k$. Le signal band-limited ainsi obtenu a été simulé pour $k \in \{20, 100\}$ pour les graphes *Ambès* et *Bx*. Les signaux band-limited pour $k = 20$ et $k = 100$ sont projetés sur le graphe d'*Ambès* Figure 4.4.

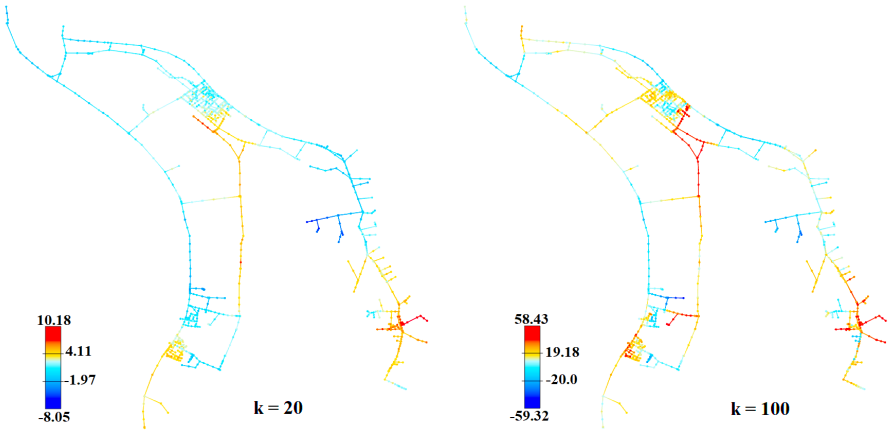


FIGURE 4.4 – Projection sur les nœuds du graphe d'*Ambès* du signal band-limited constitué à partir des $k = 20$ premiers vecteurs propres de $\boldsymbol{L}$ à gauche et $k = 100$ à droite. La couleur des arêtes est une interpolation de la couleur des nœuds adjacents pour faciliter la visualisation.

La Figure 4.5 présente les estimations $\hat{h}$ en fonction du signal band-limited Y constitué à partir de $k = 100$ vecteurs propres sur la ligne du haut et $k = 20$ sur la ligne du bas. Pour chacune des estimations un bruit $\epsilon \in \{0.1, 0.2, 0.3\}$ est ajouté au signal à reconstruire pour évaluer la capacité de reconstruction. Les points rouges représentent les nœuds ayant été échantillonnés et considérés comme observés avec dans ce cas une proportion $p = 10\%$ afin de reconstruire sur 90% des nœuds restants. Les nœuds échantillonnés sont les mêmes pour chaque estimation présentée dans le graphique. On remarque que la méthode arrive a bien reconstruire le signal et ce même avec un bruit $\epsilon = 0.3$.

De (4.1.7) on a pu exprimer $\hat{\beta} = \boldsymbol{\Phi}^T \hat{f}$ qui n'est d'autre que la représentation de la fonction f dans la base des vecteurs propres de $\boldsymbol{L}$. Les coefficients $\beta \in \mathbb{R}^{N_v}$ sont présentés Figure 4.6. Le signal à reconstruire étant band-limited et constitué à partir des $k = 20$ et $k = 100$ premiers vecteurs propres de L on remarque que l'on retrouve bien ce signal dans la base des vecteurs propres de $\boldsymbol{L}$.

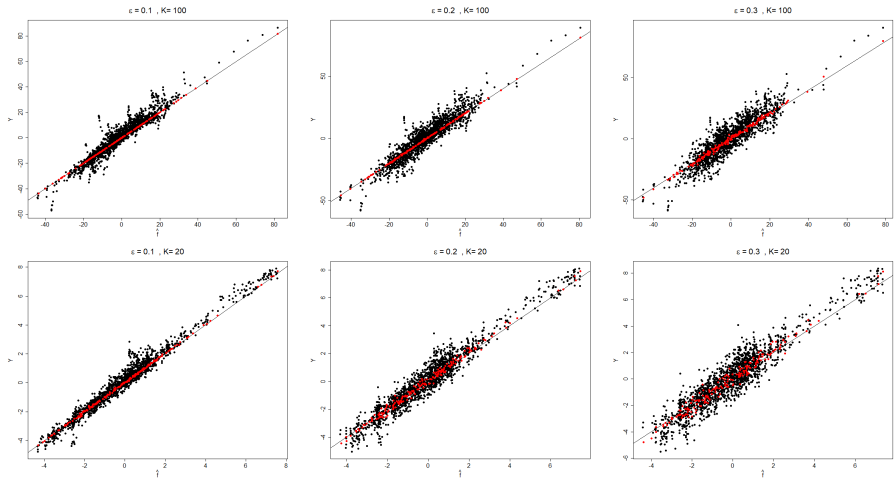


FIGURE 4.5 – Estimation d'un signal simulé band-limited, obtenu à partir de $K = 100$ vecteurs propres, resp. $K = 20$ vecteurs propres, pour la ligne du haut, resp. ligne du bas. Pour chaque colonne un bruit $\epsilon = (0.1, 0.2, 0.3)$ différent est appliqué sur le signal. En abscisse le singal reconstruit $\hat{f}$ en ordonnée le signale simulé band-limited f. Le taux d'échantillonage considéré ici est $p = 10\%$

Pour chacun des signaux band-limited testé le paramètre de pénalisation λ a été obtenu à l'aide de la méthode de validation croisée généralisée. La Figure 4.7 montre le résultat de la GCV pour l'estimation du signal band-limited avec $k = \{20, 100\}$, le niveau d'échantillonnage $p = 10\%$ et un bruit $\epsilon = 0.3$.

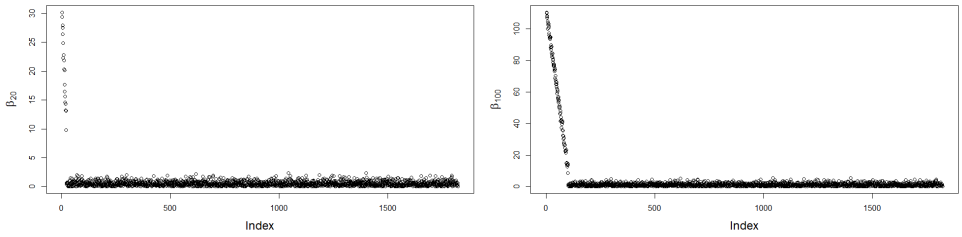


FIGURE 4.6 – Représentation de la fonction f dans la base des vecteurs propres de $\boldsymbol{L}$, dans le cas où le signal à reconstruire est une combinaison des $k = 20$, resp. $k = 100$, premiers vecteurs propres à gauche, resp. à droite.

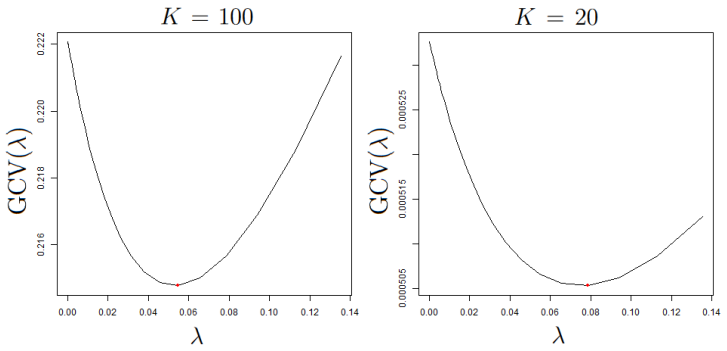


FIGURE 4.7 – Résultat $GCV(\lambda)$ pour la reconstruction du signal band-limited. $\lambda_{opt} = 0.05453$ pour le signal avec $K = 100$ vecteurs propres et $\lambda_{opt} = 0.07844$ pour $K = 20$.

4.4.3 Le signal reconstruit du débit

Dans cette section le signal que l'on cherche à reconstruire est celui des debits obtenus par la méthode de flot maximum présentée dans le Chapitre 3. Pour une journée donnée les débits sont générés sur un pas de temps de 5 minutes sur l'ensemble des nœuds des graphes d'*Ambès* et *Bx*. La Figure 4.8 présente la projection des débits estimés par la méthode du Chapitre 3 pour un temps $t = 09/12/2017$ à 13h45, sur le graphe d'*Ambès*. Une projection du signal de débits à reconstruire sur le graphe *Bx* est présentée en Annexe H.

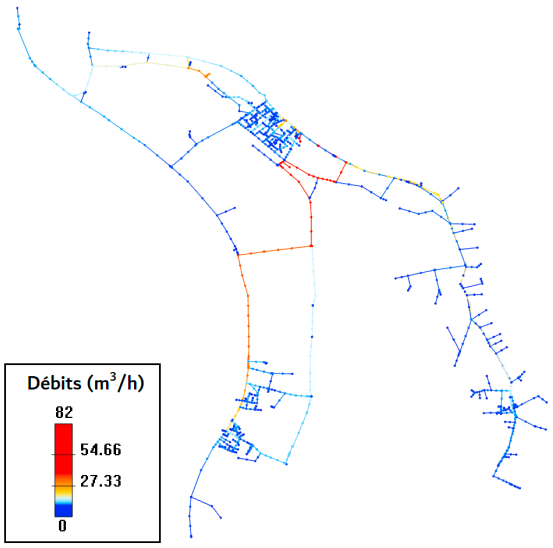


FIGURE 4.8 – Signale du débit à reconstruire, en $^{m^3}/_h$

Nous présentons l'application de la régression ridge à noyau à l'estimation des débits à partir de l'observation de ce signal sur un nombre de nœuds restreint. De la même façon que pour le signal band-limited, le signal des débits est simulé sur tous les nœuds du graphe. On est donc en mesure de tester l'efficacité de reconstruction de la méthode de régression ridge à noyau en fonction de différents échantillonnages aléatoires.

Si l'on décide de représenter les fonctions de débits f dans la base des vecteurs propres de $\boldsymbol{L}$ on s'aperçoit que le signal que l'on cherche à reconstruire n'est pas band-limited. Le signal n'est pas très "lisse" car obtenu par un algorithme d'optimisation déterministe, sans considération de régularité mais simplement d'équilibre des entrées-sorties. Au vue de la Figure 4.9 une autre interprétation pourrait être que le signal est très bruité.

Ce phénomène est d'autant plus vrai que les débits dans les canalisations sont très élevés lorsqu'il s'agit de canalisations de transport, visant à transporter l'eau d'un secteur à un autre et inversement relativement faible dans les secteurs maillés où les mouvements d'eau sont induits par les niveaux de consommation d'eau. On observe cette irrégularité sur les deux figures Annexe H et Figure 4.8 avec une majorité des nœuds avec des débits faibles (en dégradé de bleu) et des chemins complets de de canalisation avec de fortes valeurs (en dégradé de rouge).

La Figure 4.10 permet de comparer les estimations des débits obtenues par la régression ridge à noyau $\hat{f}$ en fonction du signal des débits f. Pour chacun des graphes présentés les poids de la matrice Laplacienne sont donnés par $w_{i,j} = log(\text{diam}_{i,j})$ et le paramètre λ est obtenu par la méthode GCV. Les résultats obtenus sont testés pour des niveaux d'échantillonnage différents $p = \{0.88\%, 10\%, 20\%, 60\%\}$. Les nœuds en rouge sur les graphes représentent les nœuds sélectionnés dans l'échantillonnage et dont la valeur de débits simulés a servit à la reconstruction des autres nœuds. Le paramètre de pénalisation λ_{opt} est obtenu par méthode de validation croisée généralisée pour chacun des échantillons testés. Les points en dessous de la bissectrice montrent une tendance à surestimer la

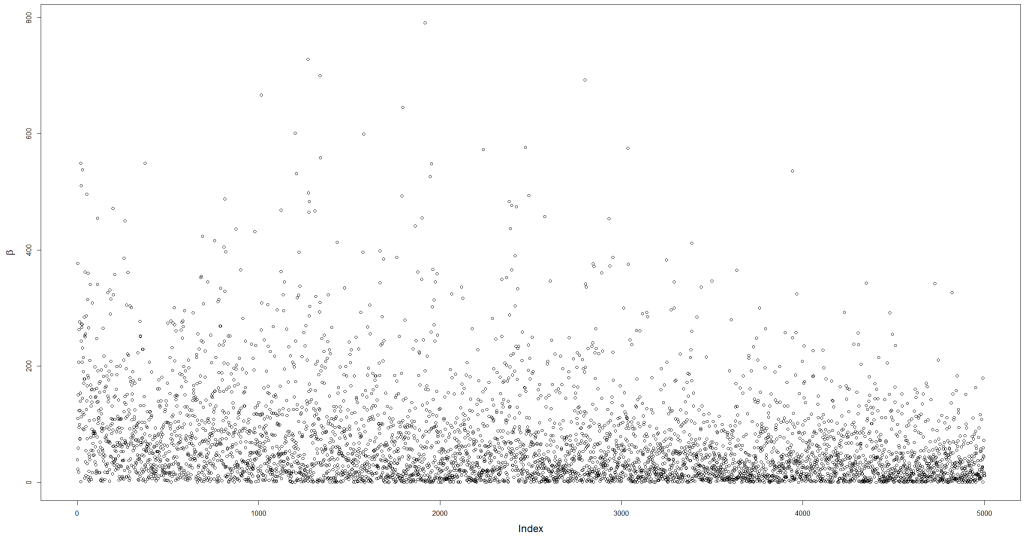


FIGURE 4.9 – Représentation des $k = 5000$ premiers coefficients $\hat{\beta}$ de la fonction des débits f dans la base des vecteurs propres de $\boldsymbol{L}$ pour la reconstruction du signal des débits sur le graphe Bx

fonction de débit lors d'estimation et inversement à sous-estimer lorsqu'ils sont en dessous. Quel que soit le niveau d'échantillonnage on observe la difficulté du modèle à reconstruire les valeurs extrêmes faisant apparaître l'irrégularité du signal sur le graphe. Certains nœuds possèdent un débit nul, il s'agit de nœuds sur en bout de réseau sur des antennes pour lesquels aucune consommation n'est mesurée, pourtant la régression ridge à noyau cherche à lisser l'information et ainsi attribue un débit sur ces antennes. Cela s'observe par les points situés en bas à gauche sur chacun des graphiques avec une ordonnée à 0 et un abscisse variant entre 0 et $15^{m^3}/_h$. Si l'on projette l'erreur résiduelle absolue $|f_i - \hat{f}_i|$ sur les nœuds du graphes Annexe I.

4.4.4 Le signal du chlore

Nous présentons dans cette partie l'application de la régression ridge à noyau à l'estimation des concentrations de chlore à partir de l'observation de cette variable sur un nombre de nœuds restreint. Le signal de chlore que l'on cherche ici à reconstruire est définit en des points précis des réseaux Bx et $Ambès$ équipés de capteurs qualités. Pour chacun de ces nœuds une série temporelle de concentration de chlore en $^{mg}/_L$ est mesurée sur un pas de temps de 15 minutes.

Le signal de chlore est alors reconstruit séparément en M fonctions $\hat{f}(t_l), l = 1, \ldots, M$, avec $M = 1000$ pas de temps. On obtient alors les résultats sur $M = 1000$ pas de temps pour fournir la prédiction $\hat{f} = (\hat{f}(t_1), \hat{f}(t_2), \ldots, \hat{f}(t_M))$. $\hat{f}$ est alors une matrice de taille $(N_v \times M)$ pour laquelle $\hat{f}_{(i,.)}$ représente le signal reconstruit pour le nœud i sur les M pas de temps et $\hat{f}_{(.,j)}$ le signal reconstruit sur tous les nœuds du graphe au temps $M = j$.

Ne disposant pas de fonction définie sur tous les nœuds du graphe il nous est impossible d'évaluer les performances de la méthode sur les signaux de concentration de chlore mesuré. Nous appliquons directement la régression ridge à noyau à partir de l'observa-

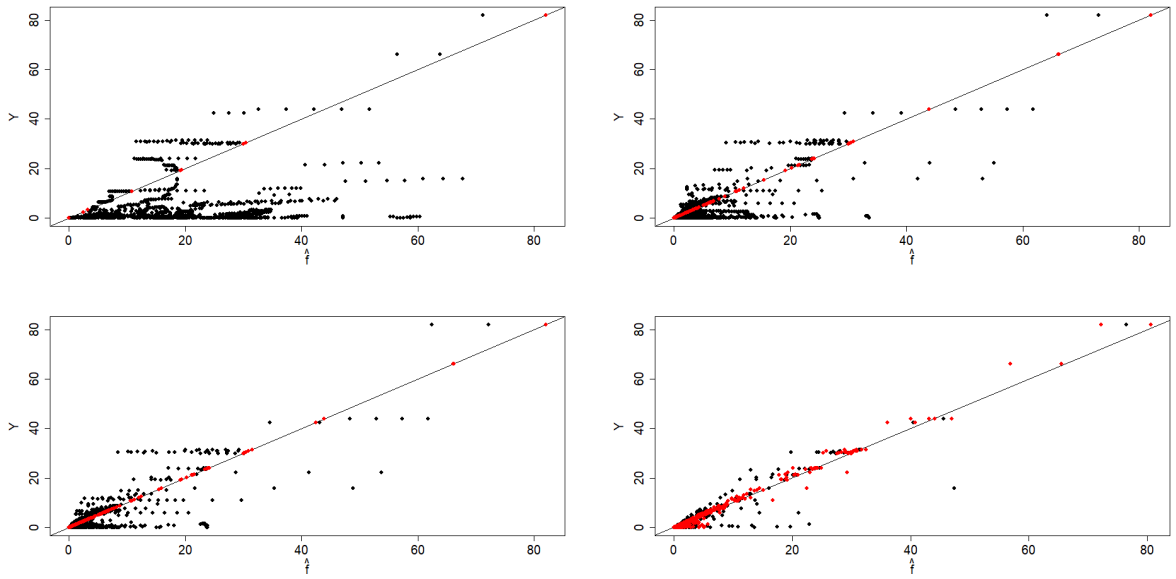


FIGURE 4.10 – Estimation des débits $\hat{f}_i$ en fonction de $Y = f_i$, $\forall i \in \{1, \ldots, N_v\}$ sur le graphe d'Ambès avec différentes proportions d'échantillonnage aléatoire, $\frac{|V_{obs}|}{N_v} \times 100 = \{0.88\%, 10\%, 20\%, 60\%\}$. En rouge les nœuds considérés comme observés $i \in V_{obs}$

tion du chlore sur le sous-ensemble de nœuds $V^{obs} \subseteq V$. Dans le cas du graphe de *Bx* $|V_{obs}| = 50$ et pour *Ambès* $|V_{obs}| = 3$.

Le signal reconstruit sur quatre des nœuds observés $u \in V_{obs}$ est présenté Figure 4.11. Pour chacun des nœuds sont présentés : en noir le signal réel mesuré par le capteur qualité au nœud u ; en vert le signal $\hat{f}_u$ représentant les concentrations de chlore reconstruites au nœud u ; et en rouge le signal reconstruit au nœud u sans prendre en compte les observations du signal de chlore du capteur u. Le paramètre de pénalisation λ est fixé.

Les signaux sélectionnés présentent des caractéristiques classiques de signaux de chlore mesurés par des capteurs qualités. En haut deux signaux bien définis avec des variabilités de concentrations de chlore différentes. Le signal A avec des concentrations de chlore élevées, proche d'une source d'eau potable est entouré de plusieurs capteurs qualités et le signal B plus isolé sur le réseau, avec des concentrations de chlore plus faibles et un signal beaucoup plus variable. Les deux signaux du bas représentent des mesures de chlore sur des capteurs disposant d'anomalies de mesures. Le signal C issu d'un capteur défaillant pour lequel la mesure de chlore dérive. Le capteur n'est plus calibré et la valeur mesurée est croissante. Et enfin le signal D qui représente un signal pour lequel le capteur mesure des piques de concentration à certains pas de temps précis et une concentration nulle en chlore par phase. Nous avons choisi de présenter ces signaux pour représenter le fait qu'au-delà du faible nombre de nœud observés la qualité des données brutes peut grandement impacter la qualité de reconstruction des concentrations de chlore que l'on souhaite obtenir.

De plus la méthode utilisée cherche à lisser l'information mesurée sur la totalité du graphe, en cherchant une certaine régularité dans sa structure. La structure du noyau prend en compte cette distance entre les observations lors de la reconstruction du signal. Néanmoins chacune des observations joue un rôle dans la reconstruction du signal et ce

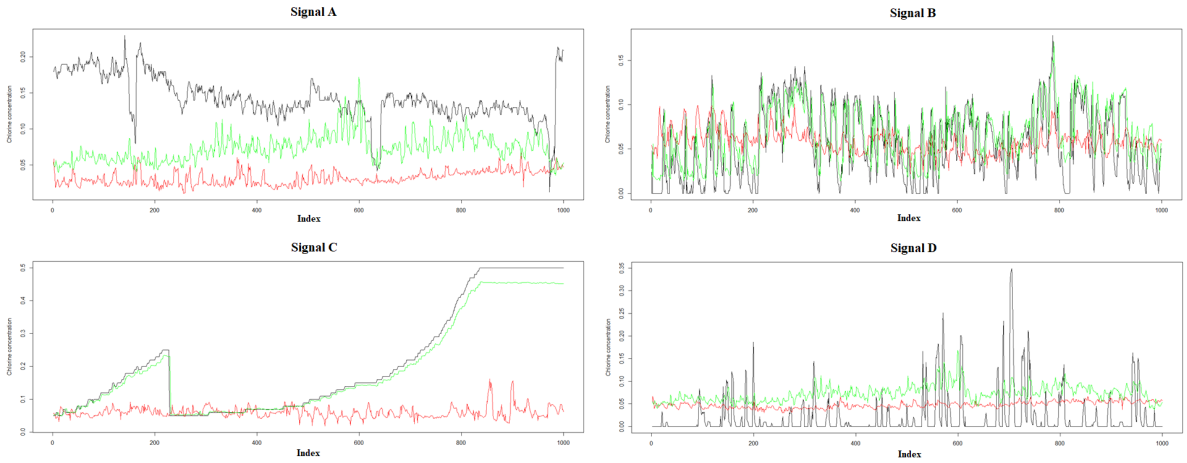


FIGURE 4.11 – Signal de chlore reconstruit par régression Ridge à noyau sur quatre nœuds $u \in V_{obs}$. En noir le signal mesuré par le capteur qualité, en vert le signal reconstruit sur le nœud u avec $u \in V_{obs}$ et en rouge lorsque $u \notin V_{obs}$

même si les deux signaux sont indépendants d'un point de vue hydraulique. En effet dans le cas d'un graphe comme celui de Bx le signal de chlore n'est pas issu d'une source d'eau unique, avec une concentration en chlore initiale décroissante. L'eau parcourant le réseau Bx provient de sources différentes, de qualités et de concentrations en chlore initiales différentes. Certaines de ces eaux se mélangeront tant dis que d'autres seront indépendantes. Ceci signifie que le chlore mesuré par deux capteurs suffisamment éloignés dans le réseau peut provenir de deux eaux complétement différentes comme l'illustre la projection de la zone d'affluence de l'usine de production d'eau potable de CapRoux en Annexe K. On comprend bien ici que la topologie structurelle du réseau ne justifie pas à elle seule les interactions possibles entre les mesures de concentrations de chlore effectuées sur le réseau.

La Figure 4.12 représente la projection du signal reconstruit de chlore $\hat{f}(t_i)$ sur tous les nœuds du graphe à partir des mesures des concentrations de chlore sur les $|V_{obs}| = 50$ capteurs (0.08% des nœuds), au temps $t_i = 13h30$ le 12 décembre 2017. Les concentrations de chlore reconstruites sont localement lisses. On détecte bien les zones où les concentrations de chlore sont les plus élevées (proches des usines de production et des réservoirs) et les zones plus éloignées où les concentrations de chlore sont plus faibles (les plus éloignés des sources). Les positions des capteurs mesurant les séries temporelles de chlore de la Figure 4.11 sont pointées sur le graphe.

Une manière de prendre en compte l'indépendance entre certains signaux serait d'effectuer la reconstruction par sous-graphes. Les sous-graphes seraient déterminés par rapport à une réalité hydraulique connue ou mesurée sur le réseau. Ainsi les signaux utilisés dans chacun des sous-graphes auraient une certaine dépendance entre eux permettant un lissage plus cohérent des valeurs observées à l'échelle du réseau. Effectuer ces calculs sur des sous-graphes permettrait aussi une réduction de dimensions des objets manipulés, permettant l'obtention des coefficients de la matrice $H(\lambda)$ qui offre un niveau d'interprétabilité supplémentaire des résultats obtenus.

Nous n'avons pas pu durant la thèse utiliser un signal de chlore simulé en tout point du

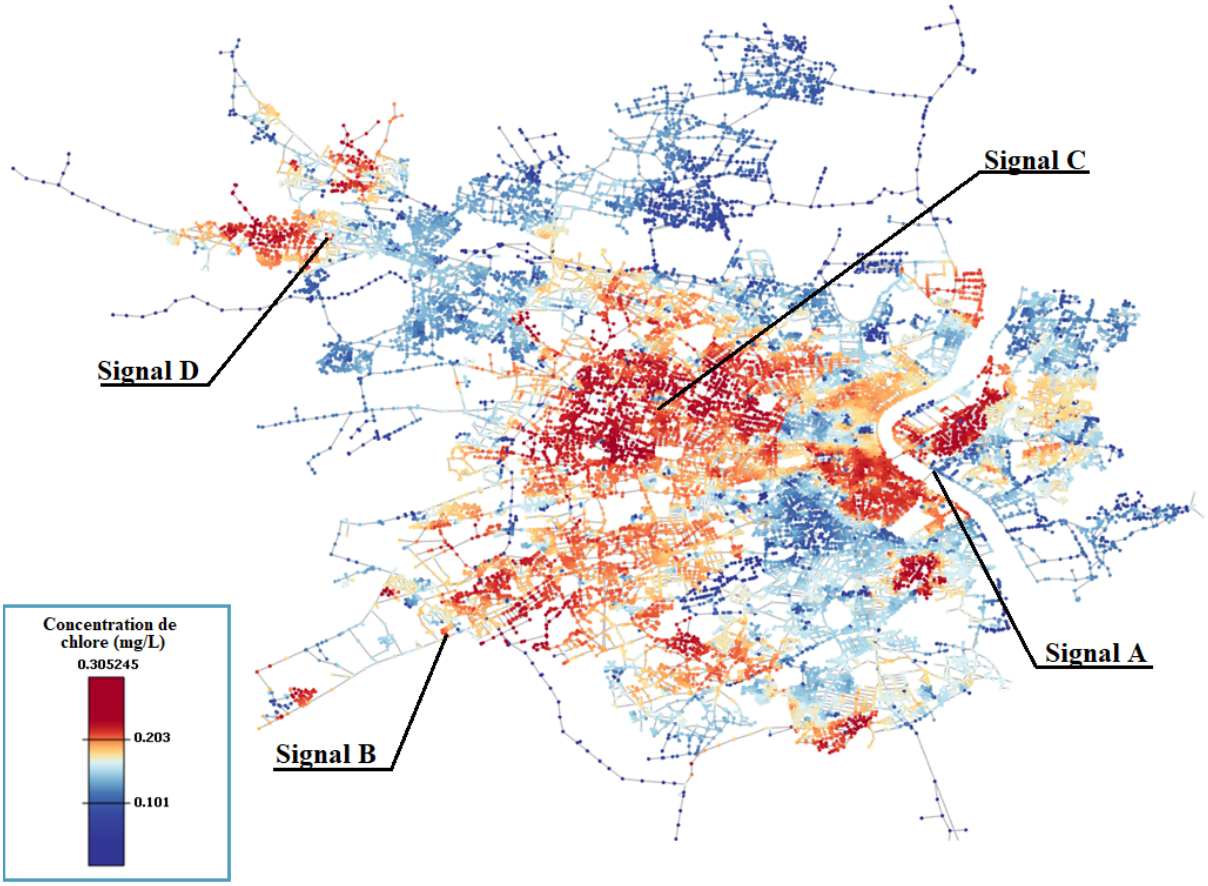


FIGURE 4.12 – Projection du signal de chlore $\hat{f}(t_i)$ au temps $t_i = 13h30$ le 12 décembre 2017, sur tous les nœuds du graphe Bx

.

réseau de distribution. Néanmoins l'obtention d'un tel signal (comme pour celui des débits du Annexe 3), à partir d'un modèle hydraulique par exemple, permettrait d'analyser la reconstruction d'un signal de chlore à l'échelle d'un réseau d'eau potable. On serait en mesure de déterminer les zones à équiper de capteur afin d'estimer la qualité de l'eau en en tout point du réseau et en temps réel à partir de l'observation du signal par les capteurs.

Conclusion

La méthode initiale, tirée de Kolaczyk et proposée ici requiert pour la définition du noyau le calcul d'une pseudo-inverse sur le graphe tout entier (60 000×60 000). Ce noyau global permet ensuite par le théorème du représentant de diminuer la taille du problème d'estimation à la recherche de n coefficients, basés sur l'extraction de $K^{(n)}$, de taille (n × n), dont la diagonalisation n'est pas lourde. Ainsi, l'effort d'inversion du Laplacien rend ensuite les calculs plus simples, la recherche du paramètre λ par validation croisée possible, et l'analyse des coefficients de reconstruction ponctuelle très intéressante (Section 4.4).

Pour une approche qualitative sur la structure en fréquence du signal sur le graphe, une recherche de valeurs propres du Laplacien est aussi très utile. Ce sont finalement les plus petites valeurs propres non nulles de L qui donne de l'information sur l'aspect lisse d'un signal sur le graphe. Évidemment, les algorithmes de recherche d'un ensemble de valeurs

propres (plutôt que le spectre entier) sont alors très utiles et on a par exemple montré que le signal de débit reconstruit par la méthode Ford-Fulkerson adaptée au Chapitre 3 renvoi un signal qui n'est pas lisse pour un Laplacien non valué, ou valué avec des poids en log-diamètre ou diamètre.

Conclusion et perspectives

La gestion d'un réseau AEP à l'échelle d'une métropole comme celle de Bordeaux génère un volume important de données et d'une grande variété toutes dépendantes de la topologie du réseau. Nous avons pu montrer qu'il était possible de gérer cette variété et volumétrie de données au sein de la structure de graphe. Cette structure offre un parallèle immédiat avec le réseau AEP. Les arêtes sont les canalisations et les noeuds les objets du réseau inter-connectant les canalisations.

La théorie des graphes nous a permis d'analyser et mieux comprendre les relations complexes existant entre chaque composant du réseau AEP. A l'aide d'algorithmes de parcours de graphes adaptés à nos besoins et une sélection des données projetées sur les noeuds et arêtes du graphe nous avons pu montrer qu'il était possible de répondre à des problématiques métiers : la planification de chantiers prioritaires à partir de données issues de modèles existants et l'ajustement des estimations de consommations par la détection automatique de secteurs hydrauliques.

Nous avons montré dans le Chapitre 3 l'adaptation de l'algorithme de flot maximum de Ford & Fulkerson à l'estimation des débits sur les noeuds d'un graphe. Cette méthodologie nous a permis à l'aide d'un réseau de flot d'estimer les sens d'écoulement et les débits à l'intérieur d'un réseau AEP à partir des données mesurées en temps réel. Ces données mesurées en certains noeuds du réseau servent de contraintes à atteindre par l'algorithme de flot, afin de déterminer une solution représentative des flux hydrauliques réellement mesurés. Les solutions fournies par l'algorithme dépendent alors fortement de la quantité et de la qualité des contraintes sélectionnées. Si la quantité est trop faible on ne peut garantir que les chemins obtenus par l'algorithme soient représentatifs des sens d'écoulement dans le réseau AEP. Si les données sont erronées ou manquantes nous ne disposons plus de points de vérifications permettant d'orienter les quantités de flux sur le réseau de flot. Pour chaque estimation l'optimisation des calculs nous permet d'obtenir une solution des débits et des sens d'écoulement en tout point du réseau en quelques secondes pour le graphe de la métropole bordelaise composé de 90 000 noeuds. Cet algorithme de flot pourrait être utilisé à des fins opérationnelles, comme par exemple pour le suivi dynamique de l'hydraulique du réseau. Cependant ce nouveau cas d'usage nécessiterait l'acquisition de données traduisant les évolutions structurelles et hydrauliques du réseau en temps réel.

Enfin dans le Chapitre 4 nous avons présenté une méthode d'inférence pénalisée sur les noeuds d'un très grand graphe représentant un réseau AEP. Nous avons utilisé une approche de prédiction par noyau reposant sur un estimateur pénalisé de type Ridge.

L'application d'une telle méthode sur le cas du réseau de la métropole bordelaise a nécessité l'inversion de très grandes matrices ce qui n'a été possible qu'avec l'utilisation de gros moyens de calcul à des fins de validation. Nous avons pu reformuler l'expression

de l'estimateur Ridge afin d'éviter l'inversion d'une grande matrice pleine et ainsi définir deux nouvelles méthodes : la méthode du représentant et la méthode itérative du gradient conjugué. Dans un cadre pratique on préférera l'emploi de la méthode itérative évitant l'inversion de grandes matrices. D'autant plus que si l'on place dans un cadre temporelle la dimension du problème vient à exploser.

L'application des trois méthodes d'estimation pour la reconstruction d'un signal simulé band-limited nous a permis de valider chacune des méthodes. Les méthodes arrivent bien à reconstruire le signal même bruité et avec un niveau d'échantillonnage faible (de l'ordre de 1% des noeuds). Les tests sur le signal des débits simulés nous ont cependant permis de constater que lorsque le signal n'est pas suffisamment régulier sur le graphe il est difficile de le reconstruire et ce même en augmentant le niveau d'échantillonnage. Enfin l'application sur le signal partiellement observé du chlore révèle que le choix du paramètre de pénalisation est crucial dans l'obtention de bons résultats lorsque le nombre de noeuds observés est trop faible. En effet, ne pouvant pas appliquer la validation croisée généralisée pour déterminer la valeur optimale du paramètre de pénalisation, nous avons pu constater un effet de moyennisation du signal sur tous les noeuds lorsqu'ils ne sont pas comptabilisés dans la base de test et que le nombre de noeuds observés est très faible par rapport au nombre de noeuds du graphe.

De plus, les différents signaux observés ne sont pas forcément réguliers sur le graphe et dépendent pour certains grandement de l'hydraulique du système (comme pour le chlore). Le noyau que l'on a choisi étant la matrice Laplacienne du graphe, elle ne prend en compte que la structure du réseau. Or dans le même graphe certains noeuds ne partagent jamais la même eau. Il peut être alors intéressant d'appliquer les méthodes de reconstruction sur des sous graphes cohérents d'un point de vue hydraulique. En plus de cela, cette approche permettrait de réduire la complexité des calculs et ainsi l'inversion de matrices Laplaciennes de faibles dimensions. On pourrait aussi envisager de tester différentes pondérations de la matrice Laplacienne ou d'autres formes de noyau prenant en compte l'aspect hydraulique sur le réseau.

Bibliographie

[Abu-Mahfouz et al., 2019] Abu-Mahfouz, A., Hamam, Y., Page, P., Adedeji, K., Anele, A., and Todini, E. (2019). Real-time dynamic hydraulic model of water distribution networks. *Water*, 11(3).

[Abu-Mahfouz et al., 2016] Abu-Mahfouz, A. M., Hamam, Y., Page, P. R., Djouani, K., and Kurien, A. (2016). Real-time dynamic hydraulic model for potable water loss reduction. *Procedia Engineering*, 154 :99–106.

[Albino et al., 2015] Albino, V., Berardi, U., and Dangelico, R. (2015). Smart cities : Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, 22 :3–21.

[Alipour et al., 2013] Alipour, Z., Monfared, M. A. S., and Zio, E. (2013). Comparing topological and reliability-based vulnerability analysis of iran power transmission network. *Proceedings of the Institution of Mechanical Engineers, Part O : Journal of Risk and Reliability*, 228 :139–151.

[Allen, 1974] Allen, D. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, page 125–127.

[Alvarez et al., 2011] Alvarez, M. A., Rosasco, L., and Lawrence, N. D. (2011). Kernels for vector-valued functions : A review. *arXiv preprint arXiv :1106.6251*.

[Angelidou, 2015] Angelidou, M. (2015). Smart cities : A conjuncture of four forces. *Cities*, 47.

[Bagler, 2008] Bagler, G. (2008). Analysis of the airport network of india as a complex weighted network. *Physica A : Statistical Mechanics and its Applications*, 387 :2972–2980.

[Barabasi and Pasfai, 2016] Barabasi, A.-L. and Pasfai, M. (2016). *Network science*. Cambridge University Press.

[Barabási, 2016] Barabási, A. (2016). *Network science*. Cambridge university press.

[Bates, 2009] Bates, D. (2009). Comparing least squares calculations. Technical report, Technical report. Available at http.

[Bavelas, 1950] Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6) :725–730.

[Belkin et al., 2004] Belkin, M., Matveeva, I., and Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory*, pages 624–638. Springer.

[Bellman, 1957] Bellman, R. (1957). Dynamic programming. *Princeton Unverisity Press*.

[Bonacich, 1987] Bonacich, P. (1987). Power and centrality : A family of measures. *American Journal of Sociology*, 92(5) :1170–1182.

[Borgatti, 2005] Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, 27(1) :55 – 71.

[Bowden et al., 2006] Bowden, G., Nixon, J., Dandy, G., Maier, H., and Holmes, M. (2006). Forecasting chlorine residuals in a water distribution system using a general regression neural network. *Mathematical and Computer Modelling*, 44 :469–484.

[Brandes, 2001] Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25(2) :163–177.

[Buckwalter, 2001] Buckwalter, D. W. (2001). Complex topology in the highway network of hungary, 1990 and 1998. *Journal of Transport Geography*, 9 :125–135.

[Caragliu et al., 2011] Caragliu, A., Del Bo, C., and Nijkamp, P. (2011). Smart cities in europe. *Journal of Urban Technology*, 18 :65–82.

[Carreras et al., 2004] Carreras, B. A., Lynch, V. E., Dobson, I., and Newman, D. E. (2004). Complex dynamics of blackouts in power transmission systems. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 14 :643–652.

[Carreras et al., 2004] Carreras, B. A., Newman, D. E., Dobson, I., and Poole, A. B. (2004). Evidence for self-organized criticality in a time series of electric power system blackouts. *IEEE Transactions on Circuits and Systems I : Regular Papers*, 51(9) :1733–1740.

[Caschili and De Montis, 2013] Caschili, S. and De Montis, A. (2013). Accessibility and complex network analysis of the u.s. commuting system. *Cities*, 30 :4–17.

[Caschilit et al., 2015] Caschilit, S., Raggiani, A., and Medda, F. (2015). Resilience and vulnerability of spatial economic networks. *Networks and Spatial Economics*, 15 :205–210.

[Chung, 1997] Chung, F. (1997). Spectral graph theory. *American Mathematical Society*.

[Cormen et al., 2009] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). Introduction to algorithms. *third edn. MIT Press, Cambridge.*

[Danladi Bello et al., 2015] Danladi Bello, A.-A., Alayande, W., Otun, J., Ismail, A., and Lawan, U. (2015). Optimization of the designed water distribution system using matlab. *International Journal of Hydraulic Engineering*, 2015 :37–44.

[Di Nardo et al., 2017] Di Nardo, A., Di Natale, M., Giudicianni, C., Greco, R., and Santonastaso, G. F. (2017). Water supply network partitioning based on weighted spectral clustering. pages 797–807.

[Di Nardo et al., 2018] Di Nardo, A., Giudicianni, C., Greco, R., Herrera, M., and Santonastaso, G. (2018). Applications of graph spectral techniques to water distribution network management. *Water*, 10 :1–16.

[Douglas Bates, 2019] Douglas Bates, M. M. (2019). *Matrix : Sparse and Dense Matrix Classes and Methods.*

[Dunn and Wilkinson, 2013] Dunn, S. and Wilkinson, S. M. (2013). Identifying critical components in infrastructure networks using network topology. *Journal of Infrastructure Systems*, 19(2) :157–165.

[Edmonds and Karp, 1972] Edmonds, J. and Karp, R. M. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM*, 19(2) :248–264.

[Elisseeff et al., 2003] Elisseeff, A., Pontil, M., et al. (2003). Leave-one-out error and stability of learning algorithms with applications. *NATO science series sub series iii computer and systems sciences*, 190 :111–130.

[Fiedler, 1973] Fiedler, M. (1973). Algebraic connectibity of graphs. *Czechoslovak Mathematical Journal*, 23(98) :298–305.

[Fortini et al., 2014] Fortini, M., Bragalli, C., and Artina, S. (2014). Identifying the high-level flow model of water distribution networks using graph theory. *Procedia Engineering*, 89 :1192 – 1199. 16th Water Distribution System Analysis Conference, WDSA2014.

[Freeman, 1977] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1) :35–41.

[Fulkerson and Ford, 1962] Fulkerson, D. R. and Ford, L. R. (1962). *Flows in Networks*. Princeton University Press.

[Giustolisi et al., 2012] Giustolisi, O., Laucelli, D., Berardi, L., and Savić, D. A. (2012). Computationally efficient modeling method for large water network analysis. *Journal of Hydraulic Engineering*, 138 :313–326.

[Godsil and Royle, 2001] Godsil, C. and Royle, G. (2001). Algebraic graph theory. *Springer*.

[Goldberg et al., 1989] Goldberg, A. V., Tardos, E., Tarjan, R. E., and Science, P. U. N. D. O. C. (1989). *Network Flow Algorithms*. Ft. Belvoir Defense Technical Information Center Apr.

[Golub et al., 1979] Golub, G., Health, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2) :215–223.

[Golub and van Loan, 2013] Golub, G. H. and van Loan, C. F. (2013). *Matrix Computations*. JHU Press, fourth edition.

[Gondran and Minoux, 1995] Gondran, M. and Minoux, M. (1995). Graphes et algorithmes. page 622.

[Guépié, 2013] Guépié, B. K. (2013). *Sequential detection of transient signals : application to the monitoring of drinking water supply network*. Theses, Université de Technologie de Troyes.

[Hager, 1989] Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM review*, 31(2) :221–239.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer Science & Business Media.

[Hawick, 2012] Hawick, K. A. (2012). Water distribution network robustness and fragmentation using graphmetrics. (762-037) :304–310. CSTN-158.

[Heineman et al., 2008] Heineman, G. T., Pollice, G., and Selkow, S. (2008). Chapter 8 : Network flow algorithms. algorithms in a nutshell.

[Herrera et al., 2016] Herrera, M., Abraham, E., and Stoianov, I. (2016). A graph-theoretic framework for assessing the resilience of sectorised water distribution networks. *Water Resour Manage 30*, pages 1685–1699.

[Horn and Johnson, 2012] Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

[Karadirek et al., 2015] Karadirek, I., Soyupak, S., and Muhammetoglu, H. (2015). Chlorine modeling in water distribution networks using arx and armax model structures. *Desalination and water treatment*.

[Kimeldorf and Wahba, 1971] Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1) :82–95.

[Kőnig, 1927] Kőnig, D. (1927). Über eine schlussweise aus dem endlichen ins unendliche. *Acta Sci. Math. (Szeged)*, pages 121–130.

[Kolaczyk, 2009] Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data : Methods and Models*. Springer Publishing Company, Incorporated, 1st edition.

[Kolaczyk and Csardi, 2014] Kolaczyk, E. D. and Csardi, G. (2014). *Statistical Analysis of Network Data with R*. Springer New York.

[Lehoucq et al., 1998] Lehoucq, R., Sorensen, D., Vu, P., and Yang, C. (1998). Arpack : Fortran subroutines for solving large scale eigenvalue problems, release 2.1. *RB Lehoucq, DC Sorensen, and C. Yang, ARPACK User's Guide : Solution of Large-Scale Eigenvalue Problems with Implicit Restarted Arnoldi Methods*.

[Lehoucq, 1995] Lehoucq, R. B. (1995). *Analysis and implementation of an implicitly restarted Arnoldi iteration*. PhD thesis.

[Liu et al., 2015] Liu, Z., Jiang, C., Wang, J., and Yu, H. (2015). The node importance in actual complex networks based on a multi-attribute ranking method. *Knowledge-Based Systems*, 84 :56–66.

[Livne and Brandt, 2012] Livne, O. E. and Brandt, A. (2012). Lean algebraic multigrid (lamg) : Fast graph laplacian linear solver. *SIAM Journal on Scientific Computing*, 34(4) :B499–B522.

[Magnanti and Orlin, 1993] Magnanti, T. L. and Orlin, J. B. (1993). *Network flows : Theory, Algorithms, and Applications*. Business & Economics, prentice hal edition.

[Mahizhnan, 1999] Mahizhnan, A. (1999). Smart cities : The singapore case. *Cities*, 16(1) :13 – 18.

[Micchelli and Pontil, 2005a] Micchelli, C. A. and Pontil, M. (2005a). *Kernels for Multi–task Learning*, pages 921–928. MIT Press.

[Micchelli and Pontil, 2005b] Micchelli, C. A. and Pontil, M. (2005b). On learning vector-valued functions. *Neural Computation*, 17 :177–204.

[Mihail, 1989] Mihail, M. (1989). Conductance and convergence of markov chains-a combinatorial treatment of expanders. *30th Annual Symposium of Foundationss of Computer Science, IEEE*, pages 526–531.

[Molitierno, 2012] Molitierno, J. J. (2012). *Applications of combinatorial matrix theory to Laplacian matrices of graphs*. CRC Press.

[Mori and Christodoulou, 2012] Mori, K. and Christodoulou, A. (2012). Review of sustainability indices and indicators : Towards a new city sustainability index (csi). *Environmental Impact Assessment Review*, 32 :94–106.

[Munavalli and Mohan Kumar, 2005] Munavalli, G. and Mohan Kumar, M. (2005). Water quality parameter estimation in a distribution system under dynamic state. *Water Research*, 39 :4287–4298.

[Nations, 2007] Nations, U. (2007). World populatoion prospects. the 2006 revision.

[Nazempour et al., 2018] Nazempour, R., Monfared, M. A. S., and Zio, E. (2018). A complex network theory approach for optimizing contamination warning sensor location in water distribution networks. *International Journal of Disaster Risk Reduction*, 30 :225–234.

[Newman, 2002] Newman, M. E. J. (2002). Assortative mixing in networks. *Phys. Rev. Lett.*, 89 :208701.

[Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2) :167–256.

[Qiu and Mei,] Qiu, Y. and Mei, J. Rspectra : Solvers for large scale eigenvalue and svd problems, 2016. *URL https ://CRAN. R-project. org/package= RSpectra. R package version 0.11-0.*

[Romero et al., 2017] Romero, D., Ioannidis, V. N., and Giannakis, G. B. (2017). Kernel-based reconstruction of space-time functions on dynamic graphs. *IEEE Journal of Selected Topics in Signal Processing*, 11(6) :856–869.

[Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*.

[Said et al., 2016] Said, A., Mourchid, M., El Faddouli, N.-e., and Mohammed, Z. (2016). Optimizing the locations of intermediate rechlorination stations in a drinking water ditribution network. *International Journal of Advanced Computer Science and Applications*, 17(12).

[Saniee Monfared et al., 2014] Saniee Monfared, M. A., Jalili, M., and Alipour, Z. (2014). Topology and vulnerability of the iranian power grid. *Physica A : Statistical Mechanics and its Applications*, 406 :24–33.

[Schölkopf et al., 2001] Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. pages 416–426.

[Scholkopf and Smola, 2001] Scholkopf, B. and Smola, A. J. (2001). Learning with kernels : support vector machines, regularization, optimization, and beyond.

[Shen and Gao, 2008] Shen, B. and Gao, Z.-Y. (2008). Dynamical properties of transportation on complex networks. *Physica A : Statistical Mechanics and its Applications*, 387 :1352–1360.

[Shuang et al., 2014] Shuang, Q., Zhang, M., and Yuan, Y. (2014). Node vulnerability of water distribution networks under cascading failures. *Reliability Engineering & System Safety*, 124 :132–141.

[Spielman and Teng, 2007] Spielman, D. and Teng, S. (2007). Spectral partitioning works : planar graphs and finite element meshes. *Linear Algebra and Its Applications*, 421.

[Spielman and Teng, 2004] Spielman, D. A. and Teng, S.-H. (2004). Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 81–90. ACM.

[Tarjan, 1972] Tarjan, R. E. (1972). Depth-first search and linear graph algorithms. *SIAM J. Comput.*, 1 :146–160.

[van Wieringen, 2015a] van Wieringen, W. N. (2015a). Lecture notes on ridge regression. *arXiv preprint arXiv :1509.09169*.

[van Wieringen, 2015b] van Wieringen, W. N. (2015b). Lecture notes on ridge regression. *arXiv : Methodology*.

[Vasconcelos et al., 1997] Vasconcelos, J., Rossman, L., Grayman, W., Boulos, P., and Clark, R. (1997). Kinetics of chlorine decay. *Journal - American Water Works Association*, 89 :54–65.

[Véron, 2006] Véron, J. (2006). *L'urbanisation du monde*. La Découverte.

[Wahba, 1990] Wahba, G. (1990). Spline models for observational data. 59.

[Wang et al., 2012] Wang, S., Hong, L., and Chen, X. (2012). Vulnerability analysis of interdependent infrastructure systems : A methodological framework. *Physica A : Statistical Mechanics and its Applications*, 391(11) :3323 – 3335.

[Yazdani and Jeffrey, 2010] Yazdani, A. and Jeffrey, P. (2010). A complex network approach to robustness and vulnerability of spatially organized water distribution networks.

[Yazdani and Jeffrey, 2011] Yazdani, A. and Jeffrey, P. (2011). Complex network analysis of water distribution systems. *Chaos : an Interdisciplinary Journal of Nonlinear Science*, 21 :016111.

[Yazdani and Jeffrey, 2012a] Yazdani, A. and Jeffrey, P. (2012a). Applying network theory to quantify the redundancy and structural robustness of water distribution systems. *Journal of Water Resources Planning and Management*, 138(2) :153–161.

[Yazdani and Jeffrey, 2012b] Yazdani, A. and Jeffrey, P. (2012b). Water distribution system vulnerability analysis using weighted and directed network models. *Water Resources Research*, 48.

[Yazdani et al., 2011] Yazdani, A., Otoo, R. A., and Jeffrey, P. (2011). Resilience enhancing expansion strategies for water distribution systems : A network theory approach. *Environmental Modelling & Software*, 26 :1574–1582.

[Yoyo et al., 2016] Yoyo, S., Page, P., Zulu, S., and A'Bear, F. (2016). Addressing water incidents by using pipe network models. *WISA Biennial 2016 Conference and Exhibition*, page 130.

[Zanella et al., 2014] Zanella, A., Bui, N., Castellani, A., Vangelista, L., and Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things Journal*, 1 :22–32.

[Zio and Golea, 2012a] Zio, E. and Golea, L. (2012a). Analyzing the topological, electrical and reliability characteristics of a power transmission system for identifying its critical elements. *Reliability Engineering System Safety*, 101 :67 – 74.

[Zio and Golea, 2012b] Zio, E. and Golea, L. (2012b). Analyzing the topological, electrical and reliability characteristics of a power transmission system for identifying its critical elements.

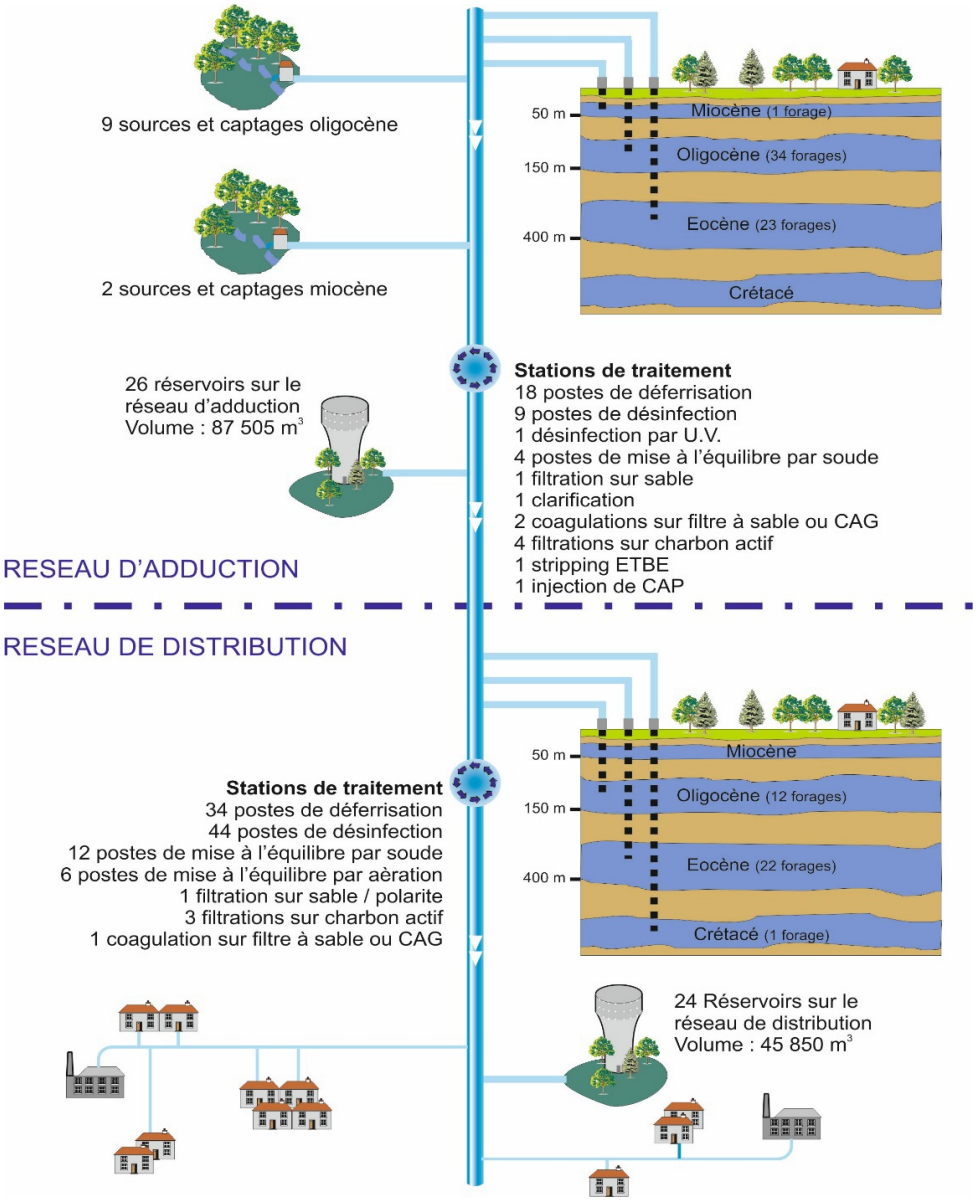
Annexe A

Communes du contrat eau de Bordeaux Métropole

Commune	Nombre d'habitants	Linéaire de réseau (Km)	Nombre de compteurs	
			TLRV	Non TLRV
Ambès	3 189	56	50	1 515
Bègles	27 589	104	432	11 425
Blanquefort	16 160	123	359	4 821
Bordeaux	253 812	648	5 320	82 743
Bouliac	3 511	40	0	1 444
Bouscat (Le)	24 037	77	340	8 397
Bruges	18 371	89	422	6 710
Cenon	24 945	87	504	6 686
Eysines	23 295	120	370	8 953
Floirac	17 142	92	427	5 850
Gradigan	25 719	144	469	7 560
Haillan (Le)	11 062	67	192	4 732
Lormont	22 690	78	483	5 928
Mérignac	71 067	329	1 530	22 227
Parempuyre	8 335	52	79	3 651
Pessac	62 260	324	943	19 571
St-Aubin-de-Médoc	7 045	72	62	2 708
St-Louis-de-Montferrand	2 237	31	31	911
St-Médard-en-Jalles	31 235	220	476	13 290
St-Vincent-de-Paul	1 032	32	10	475
Taillan-Médoc (Le)	10 147	76	106	4 327
Talence	43 506	113	551	10 560
Villenave-d'Ornon	31 967	194	487	12 492
Total	740 353	3 168	13 643	246 976

Annexe B

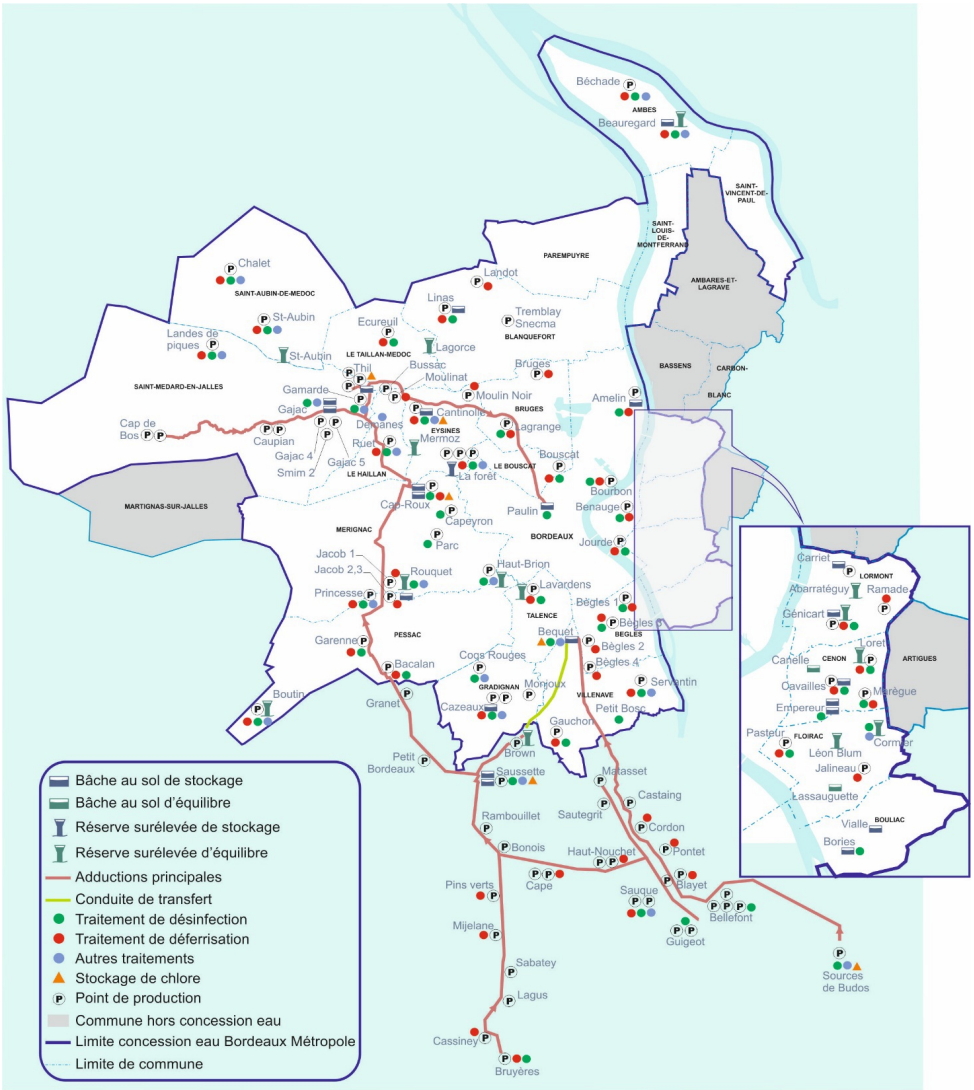
Schéma simplifié du système d'eau potable



Annexe C

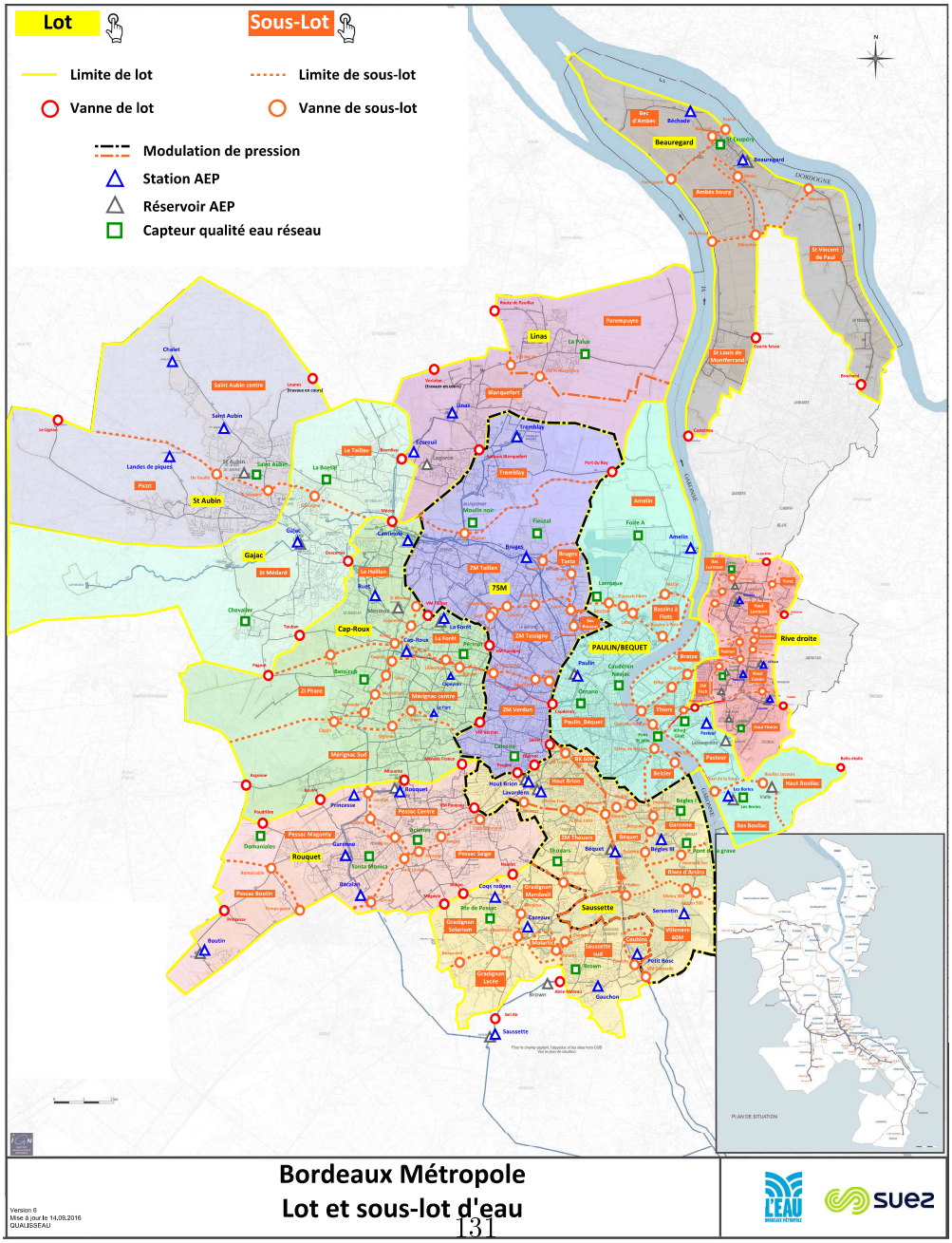
Le positionnement des installations

Le schéma et la carte figurant ci-dessous représentent la répartition des captages, des unités de traitement et des réservoirs de stockage répartis sur le réseau d'adduction et sur le réseau de distribution. Les axes principaux d'adduction et de transfert y sont également représentés.



Annexe D

Bordeaux Métropole lot et sous-lot



Annexe E

Matériaux des canalisations d'eau potable de la métropole bordelaise

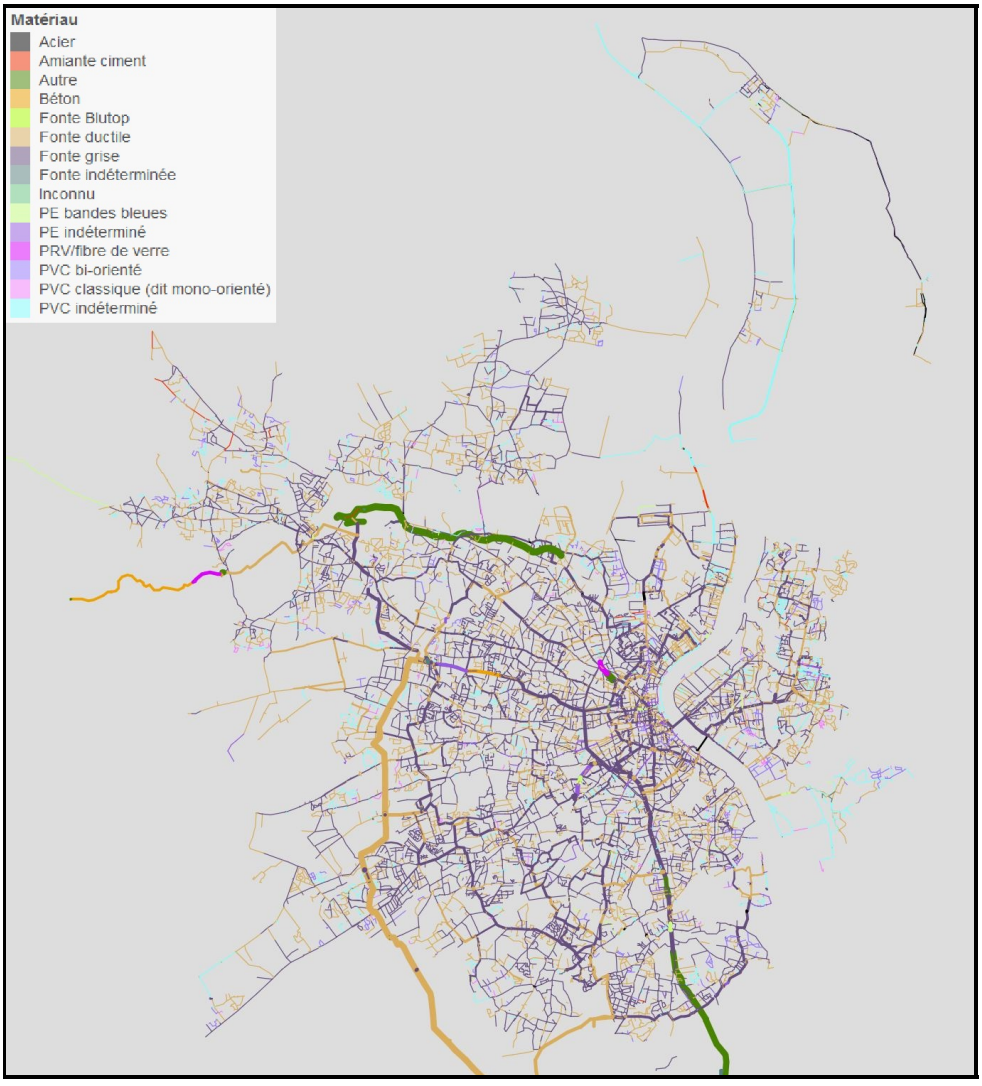
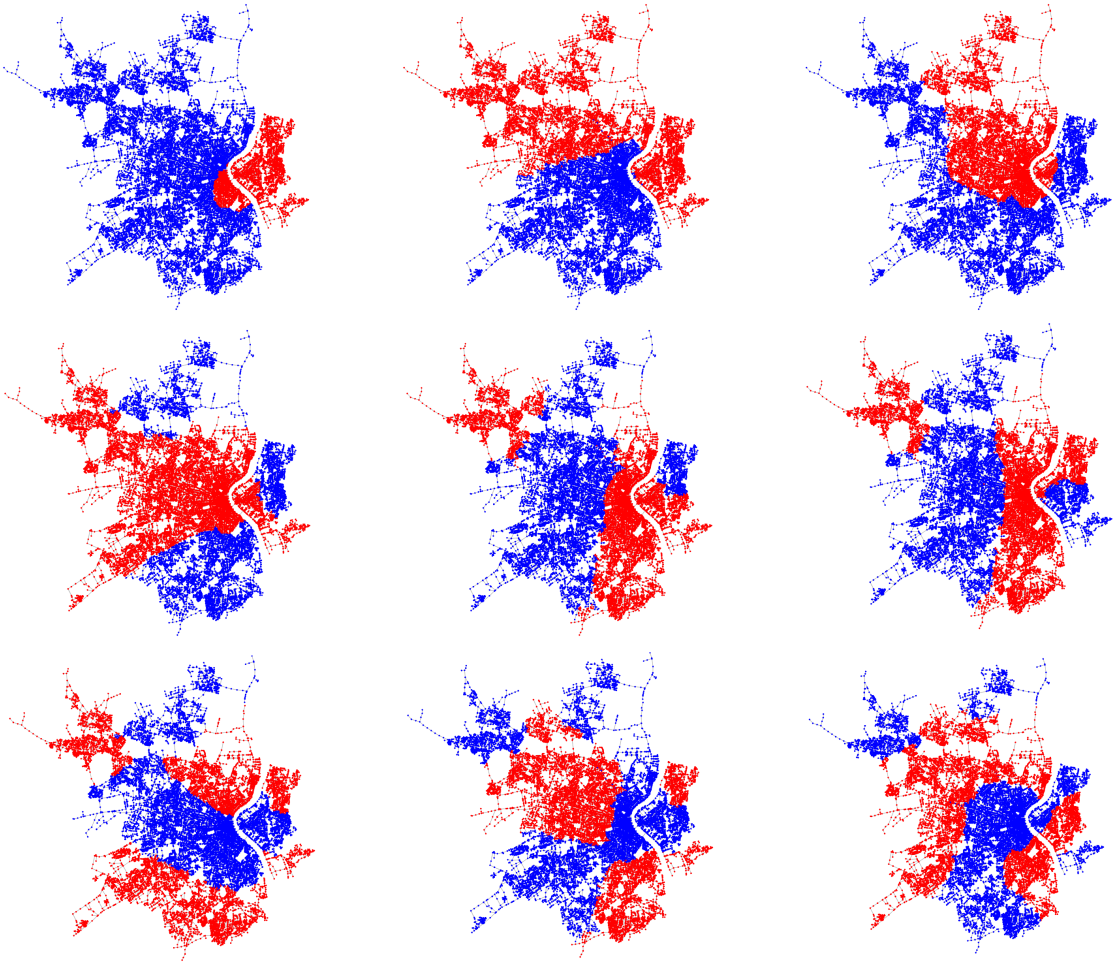


FIGURE E.1 – Conduites de distribution d'eau potable par type de matériaux

Annexe F

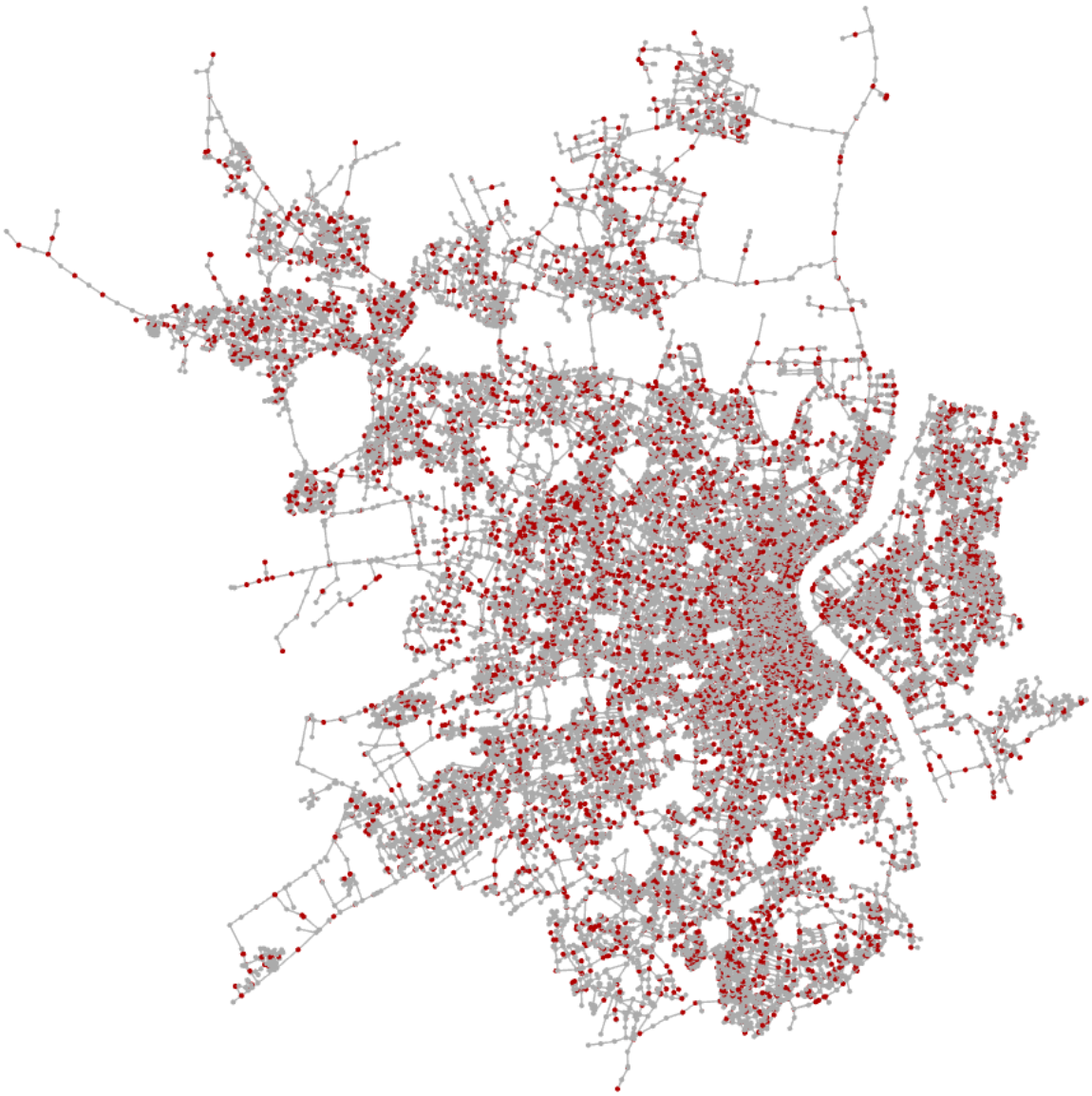
K plus petits vecteurs propres graphe *Bx*

Représentation visuelle des vecteurs propres ϕ_i associés aux 9 plus petites valeurs propres λ_i de L du réseau Bx. Rangée supérieure : $i = 2, 3, 4$; rangée du milieu : $i = 5, 6, 7$; rangée du bas : $i = 8, 9, 10$. Les valeurs négatives sont représentées en bleu et les positives en rouge.



Annexe G

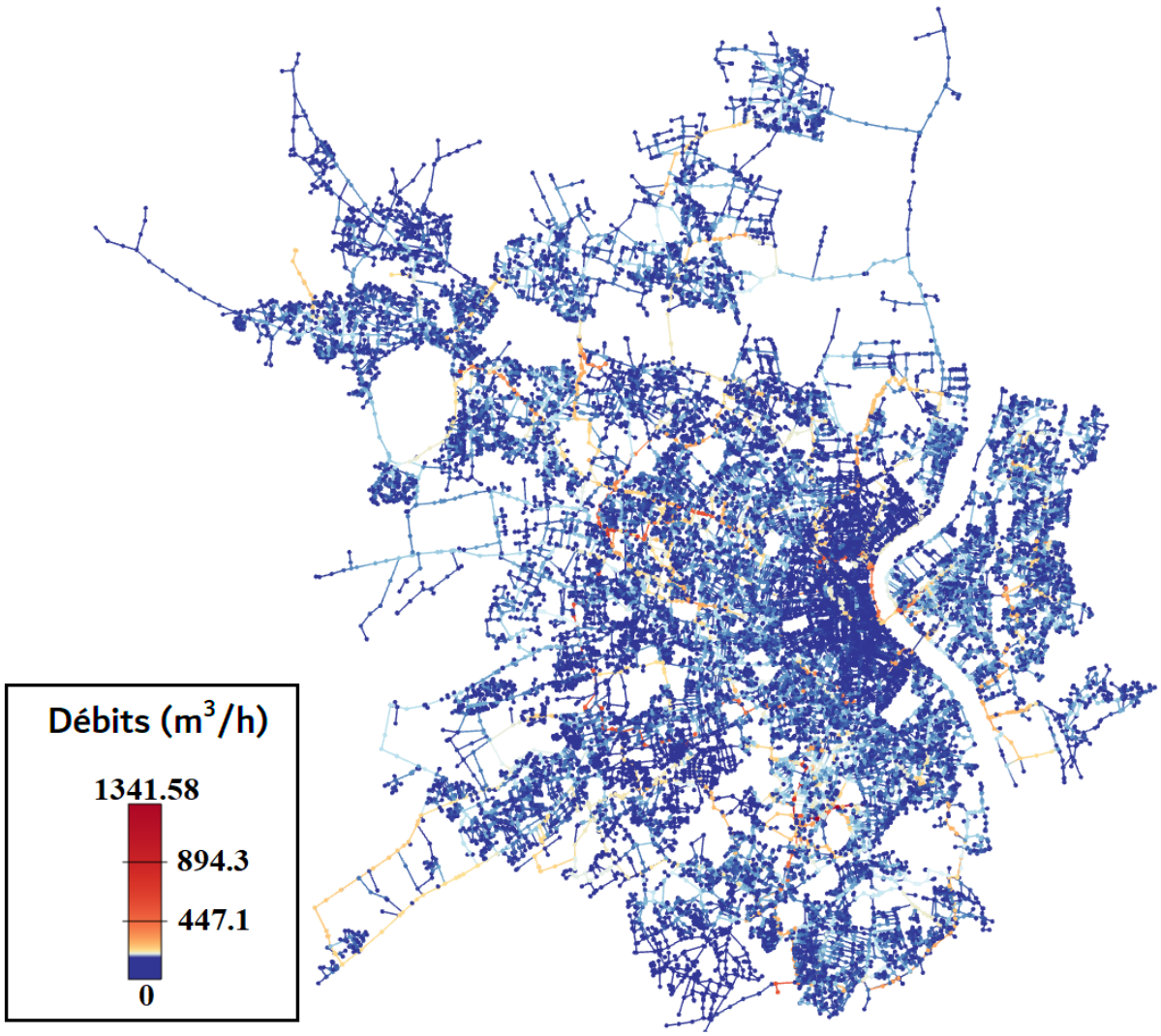
Projection des nœuds observés V_{obs} du graphe $\boldsymbol{Bx}$ $p = 20\%$



Annexe H

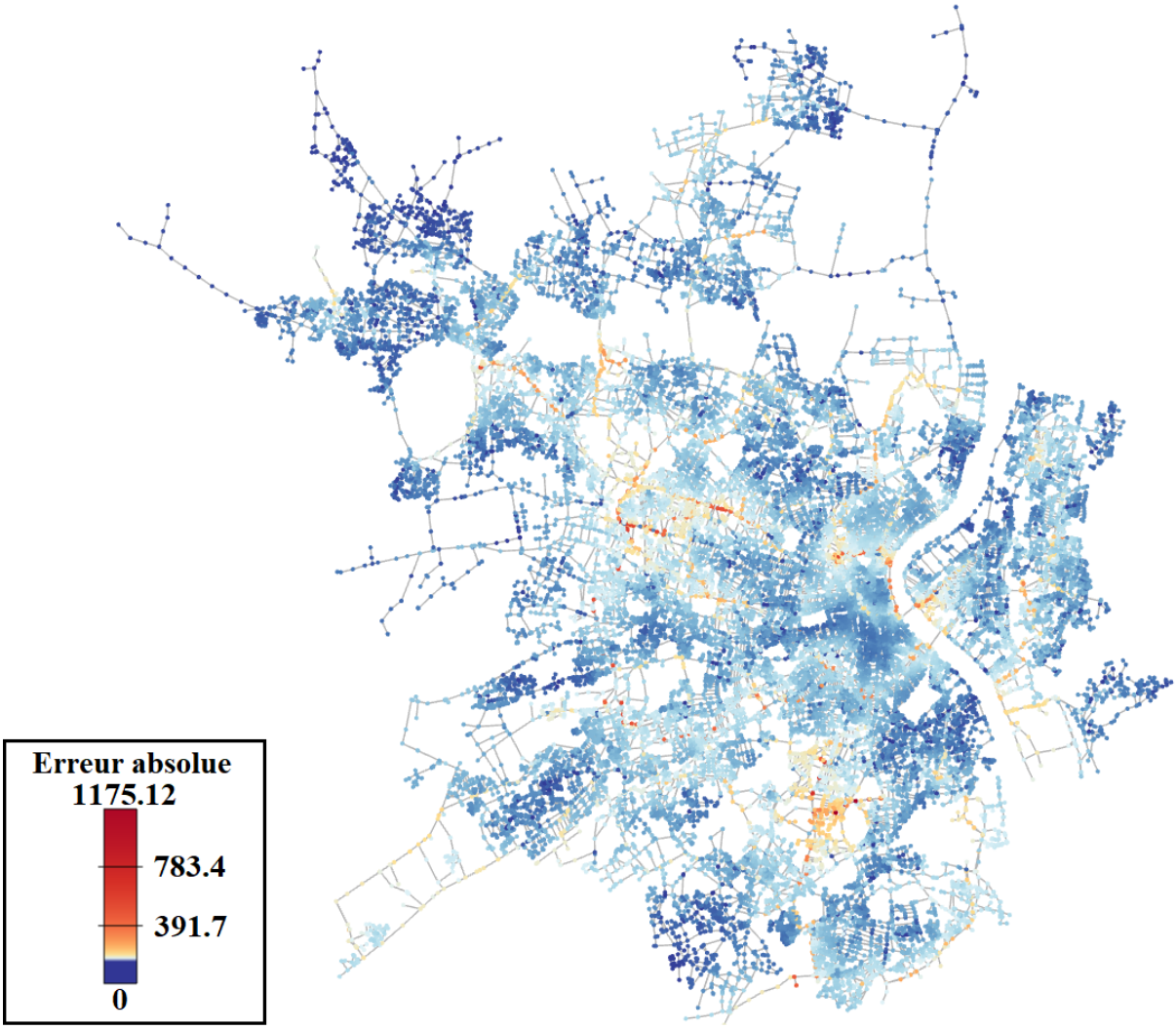
Projection des débits obtenus par la méthode de flot max du chapitre3

Le signal simulé à l'aide de l'algorithme de Ford-Fulkerson contraint est projeté sur les noeuds du graphe *Bx*. Le gradient de couleur est définit à partir du log de chaque valeur de débit afin de mieux faire apparaitre les valeurs extrêmes de débits.



Annexe I

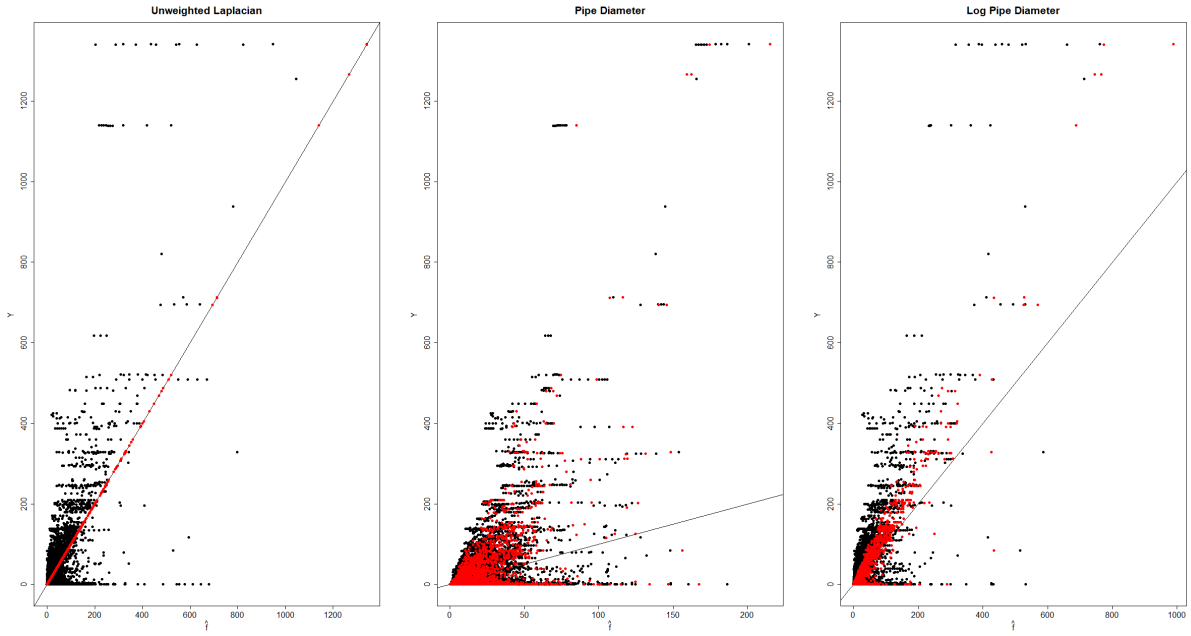
Erreur absolue de la reconstruction des débits



Annexe J

Impact des poids sur la reconstruction du signal

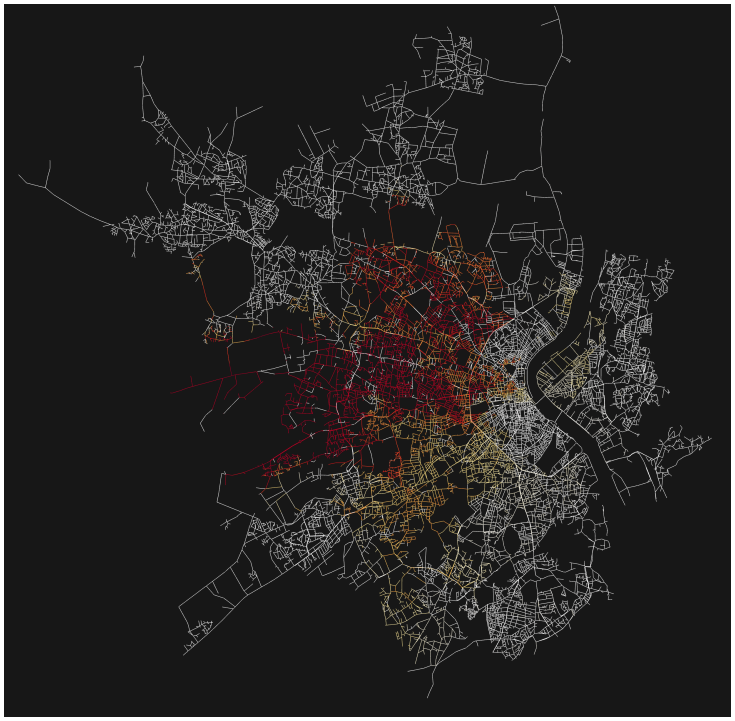
Influence des poids $w_{i,j}$ de la matrice Laplacienne du graphe sur la qualité de reconstruction du signal des débits. Le cas non podéré $w_{i,j} = 1, \forall (i,j) \in E$ 0 sinon, $w_{i,j} = \text{diam}_{(i,j)}$ et $w_{i,j} = log(\text{diam}_{(i,j)})$, où diam représente le diamètre intérieure de l'arête (i,j). Les noeuds en rouge représentent les noeuds observés $x_i \in V_{obs}$



Annexe K

Zone d'affluence d'une usine de production d'eau potable sur le réseau *Bx*

Projection sur les arêtes du graphe *Bx* de la zone d'affluence de l'usine de production d'eau potable de Cap Roux. Pour chaque arête du graphe on définit un ratio entre le débit provenant d'une usine de production d'eau potable et le débit total simulé dans la canalisation. On obtient ainsi un pourcentage pour chaque arête du graphe de 0 à 100, 100% (en rouge) signifiant que la totalité de l'eau transitant dans la canalisation provient d'une seule usine de production d'eau potable et 0% (en blanc) que l'eau issue de cette usine de production d'eau potable n'atteint pas ces canalisations. Les résultats présentés ici sont issus d'un modèle hydraulique Piccolo utilisé par Suez pour le réseau AEP de la Métropole Bordelaise.



Estimation de paramètres clés liés à la gestion d'un réseau de distribution d'eau potable : Méthode d'inférence sur les nœuds d'un graphe

Résumé :

L'essor des données générées par les capteurs et par les outils opérationnels autour de la gestion des réseaux d'alimentation en eau potable (AEP) rendent ces systèmes de plus en plus complexes et de façon générale les événements plus difficiles à appréhender. L'historique de données lié à la qualité de l'eau distribuée croisé avec la connaissance du patrimoine réseau, des données contextuelles et des paramètres temporels amène à étudier un système complexe de par sa volumétrie et l'existence d'interactions entre ces différentes données de natures diverses pouvant varier dans le temps et l'espace. L'utilisation de graphes mathématiques permet de regrouper toute cette diversité et fournit une représentation complète des réseaux AEP ainsi que les évènements pouvant y survenir ou influer sur leur bon fonctionnement. La théorie des graphes associées à ces graphes mathématiques permet une analyse structurelle et spectrale des réseaux ainsi constitués afin de répondre à des problématiques métiers concrètes et d'améliorer des processus internes existants. Ces graphes sont ensuite utilisés pour répondre au problème d'inférence sur les noeuds d'un très grand graphe à partir de l'observation partielle de quelques données sur un faible nombre de noeuds. Une approche par algorithme d'optimisation sur les graphes est utilisée pour construire une variable numérique de débit en tout noeuds du graphe (et donc en tout point du réseau physique) à l'aide d'algorithme de flots et des données issues des débitmètres réseau. Ensuite une approche de prédiction par noyau reposant sur un estimateur pénalisé de type Ridge, qui soulève des problèmes d'analyse spectrale de grande matrice creuse, permet l'inférence d'un signal observé sur un certain nombre de noeuds en tout point d'un réseau AEP.

Mots-clés : Inférence statistique, Théorie des graphes, Réseau de distribution d'eau potable, Régression Ridge à noyau, Algorithme de flots

---

Estimation of key parameters related to the management of a drinking water distribution network : Inference method on the nodes of a graph

Abstract : The rise of data generated by sensors and operational tools around water distribution network (WDN) management make these systems more and more complex and in general the events more difficult to predict. The history of data related to the quality of distributed water crossed with the knowledge of network assets, contextual data and temporal parameters lead to study a complex system due to its volume and the existence of interactions between these various type of data which may vary in time and space. This big variety of data is grouped by the use of mathematical graph and allow to represent WDN as a whole and all the events that may arise therein or influence their proper functioning. The graph theory associated with these mathematical graphs allow a structural and spectral analysis of WDN to answer to specific needs and enhance existing process. These graphs are then used to answer the probleme of inference on the nodes of large graph from the observation of data on a small number of nodes. An approach by optminisation algorithm is used to construct a variable of flow on every nodes of a graph (therefore at any point of a physical network) using flow algorithm and data measured in real time by flowmeters. Then, a kernel prediction approach based on a Ridge estimator, which raises spectral analysis problems of a large sparse matrix, allow the inference of a signal measured on specific nodes of a graph at any point of a WDN.

Keywords : Statistical inference, Graphe theory, Water distribution network, Kernel Ridge regression, Flow algorithm

---