ENSG Géomatique

ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

GRAND LYON

**Rapport de stage**

**Cycle des Ingénieurs diplômés de l'ENSG 3ème année**

# Amélioration de la connaissance patrimoniale des réseaux d'assainissement et eaux pluviales de la Métropole de Lyon

Crédit photo : Rod de Santis

Jean-Baptiste BESNIER

FONDATION ENSG - Géomatique
Sous l'égide de la Fondation de France

Le 28 septembre 2022

Non confidentiel

**Commanditaire :**
Gilles CHUZEVILLE,
Direction du Cycle de l'Eau, service Pilotage Assainissement GEMAPI Unité Gestion des Patrimoines
20, rue du lac, CS 33569, 69505, Lyon CEDEX 03, France

**Encadrement du stage :**
Gilles CHUZEVILLE, Unité Gestion des Patrimoines, maître de stage.
Cécile DUCHENE, Enseignant-Chercheur, ENSG/IGN, rapporteur principal.

**Stage** du 4/04/2022 au 16/09/2022
**Nombre de pages :** 57 et 14 d'annexes
**Version :** 2.0

| Date | Modifications |
|---|---|
| 20/06/2022 | Création |
| 01/09/2022 | Relectures |
| 05/09/2022 | Restructuration |
| 08/09/2022 | Relectures |

*Ce rapport de stage est dédicacé aux Potes Cartes, à l'équipage légendaire du Potus Cartus et à Steph qui m'ont permis de tenir le coup. Merci.*

# Remerciements

Je tiens à remercier mon maître de stage Gilles Chuzeville pour m'avoir accordé sa confiance et son soutien malgré son emploi du temps très chargé.

Je veux également remercier Frédéric Cherqui, enseignant chercheur au laboratoire DEEP, pour m'avoir suivi et conseillé tout au long de ce stage.

Je tiens à exprimer ma gratitude à Aurélie Laplanche qui m'a accueillie et nourrie dans son bureau et qui m'a grandement aidé à intégrer les us et coutumes de la métropole.

Un grand merci à tout le 3ème étage du Triangle qui a été vraiment accueillant pendant ces 6 mois de stage.

Je tiens aussi à remercier Cécile Duchêne, mon professeur référent, qui m'a suivie et très bien conseillé pendant mon stage.

Je veux également mettre en valeur l'investissement de Purdey Chevalier et Abdalah Gougui qui m'ont aidé au niveau technique pour prendre en main les données et les logiciels de la métropole.

Pour finir, je remercie mes amis et ma famille pour leur relecture.

# Sommaire

# Résumé

Ces travaux ont été réalisés dans la Direction du Cycle de l'Eau, service Pilotage Assainissement GEMAPI de la Métropole de Lyon.

L'Agence de l'Eau, via des indicateurs de connaissance des matériaux et des années de pose du réseau, accorde à la Métropole de Lyon des subventions pour gérer le patrimoine de son réseau d'assainissement. Suite au projet Hireau (Cherqui et al, 2021 **[1]**) qui propose des méthodes pour retrouver ces informations manquantes, un stage et une alternance ont déjà été réalisés. Les travaux présentés dans ce rapport ont donc pour but d'analyser, de reprendre et d'améliorer les résultats obtenus. Aujourd'hui, grâce à ces projets, la métropole remplit les critères de connaissance des années de pose et des matériaux des tuyaux constituant le réseau de collecte des eaux usées et de pluie.

Deux axes de recherches ont donc été développés pour répondre aux problématiques posées. Un premier porte sur de l'apprentissage machine avec une implémentation de l'état de l'art en termes d'estimation de données tabulaires : l'algorithme XGBoost. Ce stage a également permis la mise en place d'un processus d'optimisation de ces méthodes pour correspondre au jeu de données lyonnais. Pour les matériaux, la **précision** et le **rappel** passent de 0,59 et 0,58 à **0,89** et **0,87**. Pour les années de pose, la plage d'estimation est passée de 18 à **12 ans**.

Le deuxième axe porte sur la remontée de connaissances multi-sources. L'étude des matériaux notés dans les rapports d'inspections télévisés ont permis de récupérer **5,6 %** de données supplémentaires. La création de scripts de propagation d'information sous contrainte par parcours de graphes ont permis de récupérer **2,7 %** d'années de pose en plus. Il a également été montré qu'avec les données utilisées, il n'existait pas de lien direct entre la construction d'un bâtiment et l'âge du réseau d'assainissement situé en dessous. Ces informations remontées ont aidé à l'obtention de meilleurs résultats par apprentissage machine.

Une fois ces missions réalisées, il a été montré qu'en l'état actuel des connaissances et des données mises à disposition étudiées, il n'était pas possible, par apprentissage machine de déterminer l'état de santé des collecteurs définis par une note du programme Indigau. À la fin de ce stage, la totalité des informations des inspections télévisées ont été remontées dans trois couches SIG : couche des inspections, couche des points d'observations et couche des amorces de branchement. Ces données n'avaient jusque-là, jamais été valorisées en base de données géographique.

**Mots-clefs :** réseaux, assainissement, collectivité, matériaux, années de pose, SIG, Inspections télévisées, archives, estimations, machine learning, Graphes, Python, FME.

# Summary

This work was carried out in the Steering and Sanitation GEMAPI department of the Metropolis of Lyon.

The Water Agency, through indicators of knowledge of materials and years of installation of the network, grants the Metropolis of Lyon subsidies to manage the assets of its sewerage network. Following the Hireau project (Cherqui et al., 2021 **[1]**), which proposes methods to find this missing information, two training courses have already been carried out. The work presented in this report aims at analyzing, repeating and improving the results obtained. Today, thanks to these projects, the metropolis meets the criteria of knowledge of the years of laying and the materials of the pipes constituting the wastewater and rainwater collection network.

Two lines of research have therefore been developed to respond to the problems posed. The first one concerns machine learning with a state-of-the-art implementation in terms of tabular data estimation: the XGBoost algorithm. This internship also allowed the implementation of an optimization process of these methods to match the Lyon dataset. For the materials, the precision and recall are increased from 0.59 and 0.58 to **0.89** and **0.87**. For the years of laying, the estimation range has been reduced from 18 to **12 years**.

The second focus is on multi-source knowledge retrieval. The study of the materials noted in the cctv sewer inspections reports allowed the recovery of **5.6%** additional data. The creation of constrained information propagation scripts by graph traversal allowed the recovery of **2.7%** more years of laying. It was also shown that with the data used, there was no direct link between the construction of a building and the age of the sewer system below it. This information helped to obtain better results by machine learning.

Once these tasks were completed, it was shown that the available information held in the databases of the metropolis did not allow machine learning to determine the state of health of the collectors defined by a note from the Indigau program. At the end of this training course, all the information from the cctv sewer inspections was put into three GIS layers: the inspection layer, the observation points layer, and the branch start layer. These data had never been used in a database before.

**Key-words :** networks, sewerage, community, materials, years of installation, GIS, CCTV sewerage inspections, archives, estimations, machine learning, Graph theory, Python, FME.

# Glossaire

| | |
|---|---|
| Algorithme glouton | Un algorithme glouton implémente une méthode qui à chaque itération cherche un minimum local pour au final, espérer trouver un minimum global. |
| Arbre de décision | Un arbre de décision est un modèle d'estimation basé sous forme de questions. Graphiquement, ce modèle ressemble à un arbre où les branches sont les questions et les feuilles les résultats finaux. |
| Assainissement | Selon le guide de la gestion patrimoniale des réseaux d'assainissement conçu par l'Astee, l'assainissement est l'ensemble des techniques de collecte, de transport et de traitement des eaux usées et pluviales avant leur rejet dans le milieu naturel. |
| Benchmarking | Un benchmark est un indicateur chiffré de la performance d'une action, d'une méthode, d'un modèle, comparé à ses homologues. |
| CADEau | CADEau est le dictionnaire des données SIG de la Direction de l'Eau. Il décrit la structure des différents thèmes de données métier. |
| Collecteur | Canalisation d'assainissement fonctionnant de manière gravitaire. |
| DEEP | DEEP (Déchets Eaux Environnement Pollution) est un laboratoire de recherche de l'INSA de Lyon dont les compétences en ingénierie environnementale sont mobilisées pour répondre aux enjeux des transitions écologiques et énergétiques. |
| Effluents | Écoulement d'eau issu d'une source de pollution vers une masse d'eau naturelle, à partir d'une structure telle qu'une station de traitement des eaux usées, une conduite d'égout ou une industrie. |
| ETL | Extract-transform-load (Extraction-transformation-chargement) est une technologie permettant d'effectuer des traitements de données massifs et de réaliser des exports d'une source de donnée vers une autre. |
| F-score | Moyenne harmonique de la précision et du rappel. |
| GEMAPI | La compétence « Gestion des milieux aquatiques et prévention des inondations », plus souvent dite « Compétence GEMAPI », est en France une compétence juridique nouvelle, exclusive et obligatoire, confiée à partir du 1er janvier 2018 aux établissements publics de coopération intercommunale à fiscalité propre. Source : Wikipédia |
| Géomatique | Ensemble des outils et méthodes permettant d'acquérir, de représenter, d'analyser et d'intégrer des données géographiques. |
| GPU | Unité de traitement graphique. Certains algorithmes sont optimisés pour fonctionner dessus et ce, de manière beaucoup plus rapide que des homologues non implémentés sur GPU. |
| Graphe | En mathématiques, et plus précisément en théorie des graphes, un graphe est une structure composée d'objets dans laquelle certaines paires sont en relation. Les objets correspondent à des abstractions mathématiques et sont appelés sommets, et les relations entre sommets sont des arêtes. |
| MAPTAM | La loi du 27 janvier 2014 de modernisation de l'action publique territoriale et d'affirmation des métropoles « loi MAPTAM », est une loi française qui vise à clarifier les compétences des collectivités territoriales en créant des « conférences territoriales de l'action publique » (CTAP), organes de concertation entre les collectivités, et en réorganisant le régime juridique des intercommunalités les plus intégrées, les métropoles. Source : Wikipédia. |

| | |
|---|---|
| Métrique | Mesure. En apprentissage machine, élément permettant de qualifier un modèle. |
| Overfitting | Le sur-apprentissage, ou sur-ajustement, est une estimation qui correspond trop précisément à une collection particulière d'un ensemble de données. De ce fait, l'estimation sur d'autres données aura des résultats moindres. |
| Point fil d'eau | Cote du point le plus bas sur une zone donnée. |
| Précision | Nombre d'éléments corrects parmi ceux retournés. |
| Rappel | Nombre d'éléments correct retournés parmi ceux qui existent. |
| REGEX | Une expression régulière est en informatique, une chaîne de caractères qui décrit, selon une syntaxe précise, un ensemble de chaînes de caractères possibles. |
| RMSE | La racine de l'erreur quadratique est une mesure utilisée pour qualifier les différences entre les valeurs prédites par un modèle et les valeurs observées. |
| SIG | Un système d'information géographique ou SIG est un système d'information conçu pour recueillir, stocker, traiter, analyser, gérer et présenter de la donnée géographique. |
| Variance | la variance est une mesure qui permet de tenir compte de la dispersion de toutes les valeurs d'un ensemble de données. |

# Introduction

Dès sa mise en place en 1969, la communauté urbaine de Lyon avait la responsabilité et la compétence des réseaux d'eau potable, d'assainissement et d'eau pluviale. Elle devient au 1er janvier 2015 Métropole de Lyon. La connaissance d'infrastructures d'utilité publique est depuis lors, un des enjeux majeurs de l'unité Gestion du patrimoine. La métropole de Lyon dispose de subventions pour entretenir et améliorer son réseau. Ces subventions sont soumises à des critères d'éligibilité dont la connaissance des matériaux et les années des conduites font partie. En cela, ce papier s'inscrit dans la continuité du projet Histoire des Réseaux d'assainissement d'EAU (HIREAU) initié en 2016 qui avait pour but d'améliorer la connaissance du réseau d'eau potable et d'assainissement de la métropole (Cherqui et al, 2021 **[1]**). La fin de ces travaux a abouti à un rapport synthétique ayant le rôle de guide opérationnel pour toute collectivité qui souhaiterait reconstituer des années de pose de réseaux d'eau. Un pan de ce guide est notamment consacré à l'apport potentiel de l'informatique et donc de la géomatique dans la datation de canalisations. L'utilisation de la géomatique trouve son utilité dans de plus en plus de domaines sensibles où elle peut apporter des informations cruciales pour l'aide à la décision.

Les collectivités, grâce à leurs connaissances peuvent et doivent passer de plus en plus d'une gestion événementielle et curative à une gestion programmée et préventive. C'est pour cela que Gilles Chuzeville, responsable stratégie gestion du patrimoine des réseaux d'assainissement & eaux pluviales de la métropole de Lyon a encadré ce stage. Le but étant d'améliorer la connaissance des matériaux et des années de pose des conduites.

Dans ce rapport, seront présentés les différents axes d'approches qui ont été mis en place pour estimer et gérer les attributs[1] manquants. Premièrement, une approche par apprentissage supervisé a permis d'estimer la totalité des données avec une précision à valider. Ensuite, des algorithmes de parcours de réseau permettent de retrouver toute mesure gardée – peu d'informations – mais de manière certaine. Enfin, il a aussi été longuement discuté de l'ajout de données externes afin d'améliorer la prise de décision des algorithmes.

Il conviendra tout au long de ce rapport de valider les pistes suivies et de proposer des améliorations possibles.

Ces travaux se sont déroulés dans l'unité de gestion des patrimoines, dans le cadre du pilotage des eaux usées – eaux pluviales GEMAPI (gestion des milieux aquatiques et prévention des inondations) dans la direction du cycle eau et déchets de la métropole de Lyon. Ils sont les résultats d'un stage de troisième année d'ingénieur d'une durée de 6 mois en spécialisation Information Géographique : Analyse Spatiale et Télédétection à l'École Nationale des Sciences Géographiques.

[1] Seront appelés attributs, les informations caractéristiques d'entités géographiques. Les matériaux et les années de pose sont donc deux attributs caractérisant les éléments du réseau.

# 1. Contexte et objectifs

## 1.1. L'assainissement de la métropole de Lyon

### 1.1.1. L'organisme

En 2015, sous la loi de Modernisation de l'Action Publique Territoriale et d'Affirmation des Métropoles (MAPTAM), la communauté urbaine de Lyon devient une collectivité territoriale à statut particulier et obtient le statut de métropole. À ce titre, découlent des responsabilités dont l'objectif est d'assurer une structure politique, administrative, économique et technique à tous ses habitants.

La métropole de Lyon est un regroupement de 59 communes de plus de 1.4 million d'habitants répartis sur 534 $km^2$. Ces chiffres imposent la mise en place d'outils de gestion et de maîtrise d'aménagements du territoire et plus particulièrement, des compétences dans la maîtrise de l'eau et de l'assainissement. La métropole de Lyon conserve ses compétences liées à l'eau qui étaient déjà présentent dans la communauté urbaine. En 2018, elle se voit confier une compétence supplémentaire de Gestion des Milieux Aquatiques et Prévention des Inondations (GEMAPI).

Dans le cadre de l'eau, la métropole a donc la responsabilité du cycle de celle-ci. La direction du cycle de l'eau est composée de divers services regroupant 650 agents dont le rôle est d'assurer la gestion du réseau. Elle est propriétaire des infrastructures. C'est donc elle qui définit les prix, la stratégie, gère le patrimoine et qui programme les travaux sur les installations. Dans la partie collecte des eaux usées, l'assainissement est effectué en régie directe par le service Pilotage Assainissement GEMAPI (PAG) de 60 agents. Gestion allant de la sensibilisation du public aux enjeux de l'assainissement à la planification et l'exécution de travaux. La métropole, organisation dense et étendue, dispose donc de moyens d'assurer la gestion de son assainissement.

### 1.1.2. Système d'assainissement

L'assainissement est le procédé qui consiste à collecter, traiter et évacuer les eaux usées et pluviales produites par les particuliers, les zones d'industries et les zones urbanisées. Le but étant de rejeter ces eaux traitées en milieu naturel. Pour mener à bien cet assainissement, un ensemble de tuyaux aux caractéristiques multiples, appelés collecteurs, forment un réseau souterrain.

À la différence d'un réseau comme la distribution en eau potable ou le chauffage d'une maison, l'eau ne circule pas sous pression, mais de manière gravitaire. Chaque tronçon de tuyau possède un amont et un aval.

Trois types de réseaux sont distingués en fonction du type d'effluent transitant à l'intérieur : les eaux usées (EU), rejets domestiques ou eaux polluées par des polluants physiques, chimiques ou biologiques par un usage humain. Les eaux pluviales (EP) qui résultent d'eaux de ruissellement et les eaux du réseau unitaire (UN) qui est parcouru par un mélange d'eaux usées et d'eaux pluviales. La gestion de ces trois types d'effluent est différente avant un retour au milieu naturel et il existe pléthore d'infrastructures pour nettoyer ces eaux.

Suivant la loi Barnier de 1995, les métropoles ont obligation de publier de manière annuelle un rapport présentant notamment les caractéristiques du réseau. Ainsi, pour l'année 2020, le réseau de la Métropole de Lyon était constitué de :

- 12 stations de traitement des eaux usées (390 000 $m^3$ / jours)
- 408 déversoirs d'orages
- 222 désableurs/déhuileurs
- 300 + bassins de retenue ou d'infiltration des eaux pluviales
- 1 765 km de réseau unitaire (UN)
- 1 451 km de réseau séparatif (926 km EU | 525 km EP)

Au vu de ces informations, un programme de gestion patrimoniale apparaît comme indispensable de par la complexité et l'importance publique de la structure supportant l'assainissement. Il permet d'assurer un suivi fiable et régulier des équipements et de gérer en fonction des enveloppes budgétaires le dimensionnement et le pilotage des ouvrages. Il permet également d'agir en prévention pour préserver la santé humaine et les milieux aquatiques. La figure 1 représente le schéma complet de collecte et de traitement de ces eaux.

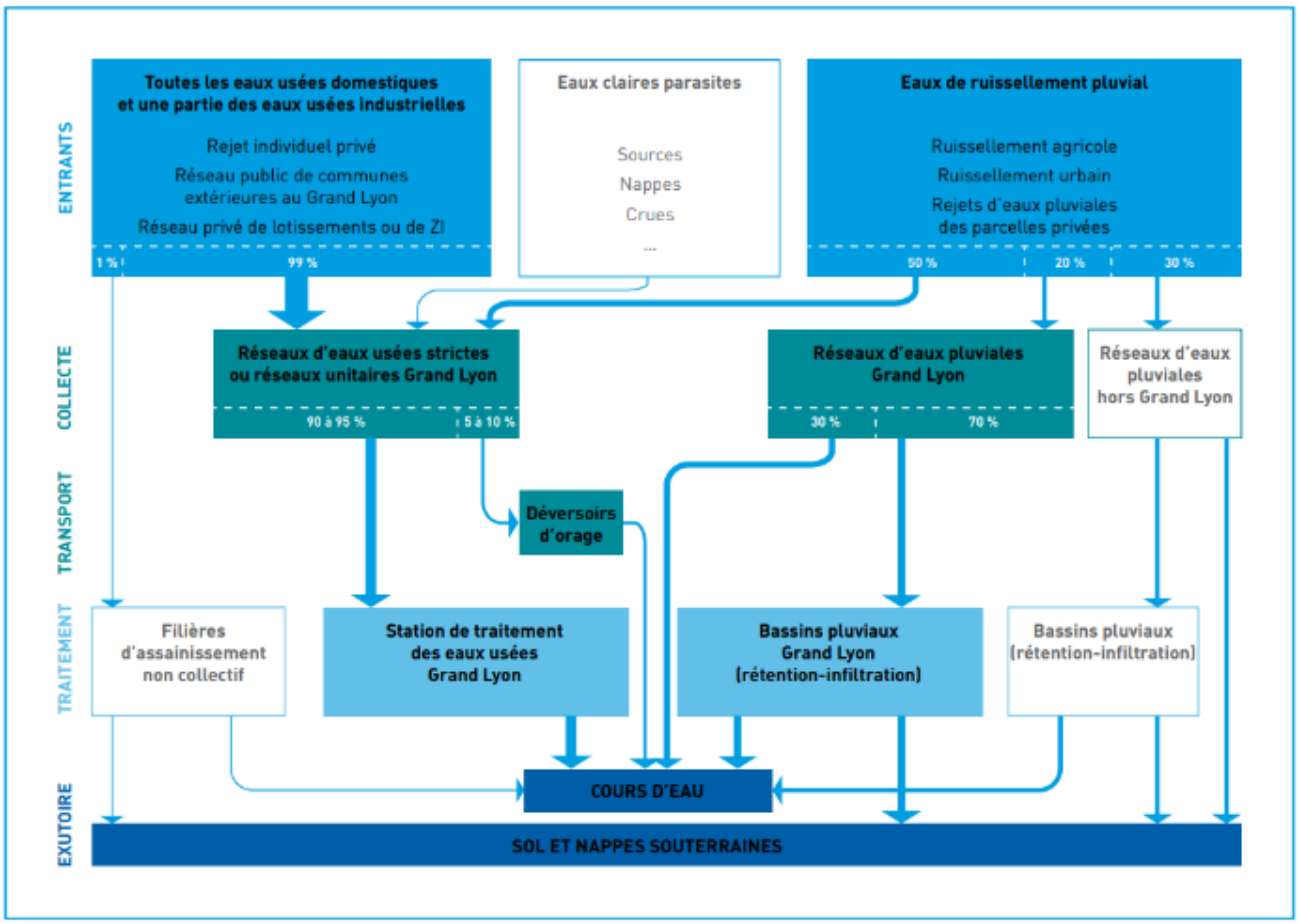



*Figure 1 Cheminement des eaux usées et pluviales dans le réseau d'assainissement métropolitain, extrait du Schéma général d'assainissement du Grand Lyon 2015-2027* **[2]**

## 1.2. Le stage

### 1.2.1. Besoin et contexte

L'observatoire national des services d'eau et d'assainissement SISPEA régi par le service public d'information sur l'eau Eau France a édité en 2013 une fiche d'indices de connaissance et de gestion patrimoniale des réseaux de collecte des eaux usées (P202.B). Cette fiche, dont un extrait est disponible en **annexe C** (p. 76), donne des points à cumuler suivant différents paliers de connaissances ou de compétences acquises. Elle est la référence en termes de gestion patrimoniale mais donne également accès à des subventions pour le maintien et l'amélioration du réseau suivant les paliers atteints.

Cette fiche P202.B n'est pas la seule à conditionner l'obtention de subventions, mais permet d'obtenir des objectifs clairs pour la métropole. Les trois critères à respecter pour l'obtention de cette subvention sont le prix de l'eau, le remplissage de l'observatoire des services et 30 points à cumuler sur cette fiche pour la période 2021-2022[2] en assainissement. Le non-respect d'un de ces critères amène à la rupture du contrat avec l'agence de l'eau Rhône Méditerranée Corse, représentant une perte minimale de 23,7 millions d'euros de subventions sur le contrat 2022-2024. L'accès à la connaissance permet la prévision et l'anticipation des opérations à effectuer sur le réseau. Cela permettrait d'éviter les travaux dit curatifs pour gérer différents types d'événement comme des fuites, des infiltrations, des effondrements, de la pollution et à fortiori des nuisances pour la population.

Pour l'instant, la métropole n'a pas débloqué les indicateurs concernant les matériaux et les années de pose. En juin 2020, 54 % des matériaux et seulement 30 % des années de pose étaient connus. Au 22 avril 2022, les données connues passent à 59 % et 31 %. Ces deux informations rapportent 0 points sur un total de 30[3]. Les matériaux ont pu être récupérés via des inspections par les équipes techniques, mais les années de pose ne peuvent être récupérées que via un travail d'archive. Il y a donc un intérêt crucial à utiliser la géomatique pour réussir à estimer les paramètres manquants des collecteurs.

Les travaux précédents se sont focalisés sur les années de pose. Ici, en ajoutant l'estimation des matériaux, les estimations sur les années de pose pourraient être améliorées et débloquer des indicateurs d'états de santé du réseau. Ces informations de matériaux peuvent également amener à la discussion autour d'une politique de désamiantage ; amiante encore utilisé jusque dans les années 2000 pour l'écoulement des stocks.

Gilles Chuzeville, dans l'Unité Gestion Des Patrimoines (GDP) a proposé de travailler sur l'amélioration de la connaissance patrimoniale des réseaux d'assainissement et eaux pluviales. Cette Unité GDP, a à charge d'améliorer la connaissance du réseau et de connaitre son état de santé afin de mettre en place un plan d'action de gestion sur la durée. Elle a également le rôle de maître d'ouvrage, c'est elle qui est l'entité porteuse du besoin qui sera réalisé par un maître d'œuvre : maîtrise d'œuvre interne ou société de travaux extérieure à la métropole.

Ce stage fait suite aux travaux d'un master 2 sur l'année 2019-2020 en stage de fin d'étude et d'un alternant en DUT sur l'année 2020-2021. Il y a donc lieu de reprendre, d'analyser et enfin d'améliorer les travaux précédemment réalisés.

[2] 60 points pour la période 2023-2024

[3] 10 points sont actuellement acquis grâce à l'existence d'un inventaire des réseaux. 10/30 avec objectif 30 fin 2022.

### 1.2.2. Objectifs

Je devais réaliser un stage de fin d'étude pour clore mon cursus au sein de l'ENSG. J'ai cherché un stage qui pouvait être complémentaire de ma formation, sur les questions d'environnement et des points non abordés lors de mes cours comme l'eau, dans le but d'avoir une vision plus globale de la géomatique.

Le service PAG a subi ces deux dernières années une restructuration qui a eu pour conséquence de saturer les demandes liées à ce service. Gilles Chuzeville étant très occupé, il a fallu malgré le suivi régulier, faire preuve d'autonomie et être force de proposition. Quatre axes de recherches pour ce stage sont apparus :

- Proposer des méthodes afin d'estimer les matériaux et les années de pose manquants de la couche SIG des collecteurs. Cet objectif, s'est scindé en deux au cours du stage : premièrement, produire un algorithme d'estimation des attributs via machine learning, deuxièmement, réussir à trouver des données supplémentaires fiables via d'autres algorithmes, afin d'alimenter davantage le premier pour améliorer les résultats.
- Proposer des méthodes en « presse-bouton » ou assez simples à débugger pour les personnes qui prendront la suite de ces travaux. De cet objectif, découlent des contraintes : travailler avec le logiciel FME (Feature Manipulation Engine) et dans le langage Python, langage le plus maîtrisé par l'équipe géomatique.
- Écrire un article pour la revue technique TSM (Techniques Sciences Méthodes, la revue mensuelle des spécialistes de l'environnement, éditée par l'ASTEE) présentant les avancées et les méthodes réalisées au cours de ce stage. Ces travaux sont effets sensibles d'intéresser d'autres collectivités.
- Remonter dans leur base de données, des informations résultant des inspections de conduites non visitables. Cet objectif se détache des trois premiers puisqu'il est arrivé en dernier après-réalisation du reste.

### 1.2.3. Démarche

Le stage s'est découpé en plusieurs grandes phases. Une première phase de récupération et de compréhension des travaux antérieurs ont permis de se rendre compte d'améliorations possibles et d'étudier l'état de l'art. Cette partie de début du stage a aussi été l'occasion de découvrir le milieu professionnel public et d'apprendre beaucoup sur les réseaux. S'est ensuite déroulée une phase de développement d'algorithmes de machine learning afin d'estimer les attributs manquants. La conclusion de cette phase a amené la question de l'ajout de données afin d'améliorer les résultats. Il faut les rendre les plus cohérents possibles compte tenu de l'expérience des agents de la métropole. C'est ainsi qu'a été développé un algorithme de propagation d'information attributaire par parcours de graphe et qu'une étude de données de datation de lotissements a été effectuée. Ensuite, les informations sur les matériaux venant des inspections des canalisations non-visitables ont été remontées. Par la même occasion, toutes les informations de ces inspections ont été mises en base de données. Enfin, ces travaux se sont terminés par une phase de rédaction.

Au niveau de la gestion du stage, pour bien tenir compte des avancées et des problèmes rencontrés, un journal de bord a été régulièrement complété ainsi qu'un tableau de bord avec l'outil Trello. Le déroulement du stage via un diagramme de Gantt est présent en **annexe H** (p. 85).

# 2. Estimation d'informations attributaire par apprentissage

## 2.1. Présentation des données

### 2.1.1. Le réseau d'assainissement et ses caractéristiques

La direction de l'eau dispose depuis 1989 d'un Système d'Information Géographique (SIG) où la donnée, stockée sur serveur est accessible sur les postes autorisés. Le SIG est structuré en cinq grands thèmes de données et comporte 154 couches. La métropole dispose d'un catalogue de donnée CADEau qui permet d'explorer les attributs et les contraintes sur les données. Dans le thème assainissement (**annexe E** p. 79), c'est la couche ASSCANAP qui porte l'information de tous les collecteurs d'assainissement. C'est une donnée linéaire géoréférencée comportant 40 attributs par entité. Sur ces 40 attributs, tous n'ont pas une complétude à 100 % et tous ne sont pas utiles pour déduire des informations. Les collecteurs sont découpés en tronçons en fonction des intersections du réseau.

Parmi ces attributs, 19 ont été sélectionnés afin de répondre à la question des années de pose et des matériaux. Ils viennent caractériser les tronçons. Les 21 non-sélectionnés sont le plus souvent des données administratives, des noms, des commentaires ou alors l'attribut est très faiblement renseigné.

- IID_IDENTCANAP : Identifiant unique du collecteur
- SCD_TYPECANAP : Type de collecteur (Principal, refoulement, siphon)
- FLT_ZAMONT / FLT_ZAVAL : Altitude en amont et en aval du collecteur
- ICD_PREXY / ICD_PREZ : Classe de précision planimétrique et altimétrique
- FLT_PENTE : Pente du collecteur (m/m)
- FLT_LONGUEUR : Longueur réelle du collecteur
- SCD_TYPERES : Type d'effluent pouvant être raccordé au collecteur (UN, EP, EU)
- INT_ANNEEREHAB : Année de réhabilitation du collecteur
- INT_ANNEEPOSE : Année de pose du collecteur
- ICD_DOMANIALITE : Domanialité du collecteur (commune, privé, SNCF, etc.)
- SCD_ENTRETIEN : Responsable entretien (Service voirie, eau, etc.)
- SCD_MATERIAU : Code du matériau du tronçon.
- ICD_STRUCT : Booléen sur le fait que le tronçon soit structurant ou non
- SID_ASSCONDUITE : Codification de la forme de la conduite
- IID_ASSSANTEHISTO : Code état de santé historique
- IID_ASSANTENV : code état de santé non-visitable
- SCD_TYPEEFFLUENT : qualification de l'eau traversant le tronçon (EU, EP, UN)
- INT_LARGEUR : Largeur du tronçon
- INT_HAUTEUR : Hauteur du tronçon

Concernant les données à produire, il existe déjà dans la couche ASSCANAP trois attributs concernant les années de pose estimées : INT_ANNPO_1, INT_ANNPOSEINF et INT_ANNPOSESUP qui correspondent aux années estimées par les travaux précédents avec les bornes Inf. et Sup. liées à la précision de l'estimation.

### 2.1.2. Regroupement des classes de matériaux

La classe des matériaux SCD_MATERIAUX est définie par 20 codes différents (Tableau 1) afin de préciser la connaissance des matériaux d'un tronçon de collecteur.

| Code | Valeur domaine | Code | Valeur domaine |
|---|---|---|---|
| ACIE | Acier | FONT | Fonte |
| AMCI | Amiante ciment | GRES | Grès |
| BATO | Béton âme tôle | MACO | Maçonnerie |
| BCAR | Béton centrifugé armé | PEHD | Polyéthylène Haute Densité |
| BCFA | Béton coulé en fouille armé | PETH | Polyéthylène |
| BCFN | Béton coulé en fouille non armé | PP | Polypropylène |
| BLPB | Béton LPB | PRV | Polyester Renforcé de fibres de Verre |
| ENDU | Enduit | PVCL | P.V.C. |
| FCAL | Fonte calorifugée | AUTR | Autre |
| FDUC | Fonte ductile | | |

*Tableau 1 Classes originelles des matériaux de la couche ASSCANAP*

Étant donné que l'estimation des matériaux se fera de manière statistique, il apparaît comme utile de regrouper les classes avec le plus de ressemblance pour obtenir les meilleurs résultats (Tableau 2). Après discussion avec les membres de l'équipe, sept classes ont été définies pour catégoriser les matériaux. Elle a été faite en prenant en compte l'occurrence des classes, la dégradation des matériaux et leur similarité.

| Description | Code | Ancien(s) Code(s) | Support |
|---|---|---|---|
| Autre | AUTR | ACIE / PLOM / AUTR | 4% |
| Béton avec métaux | BTAM | BATO / BCAR / BCFA | 61% |
| Béton autre | BTAU | BCFN / BIND / BLPB | 13% |
| Fonte | FON | FCAL / FDUC / FONT | 1% |
| Plastique | PLAS | PEHD / PETH / PP / PRV / PVCL | 17% |
| Roche | ROCH | ENDU / GRES / MACO | 3% |
| Amiante cimenté | AMCI | AMCI | 1% |
| Inconnu | Null | Null | |

*Tableau 2 Regroupement des matériaux selon leur similitude*

## 2.2. Travaux précédents

Le projet Hireau (Cherqui et al, 2021 **[1]**) a permis de synthétiser les méthodes envisageables pour aider à la reconstitution des dates de pose d'un réseau d'assainissement. Une de ces approches est la datation basée sur l'analyse via apprentissage statistique sur les données existantes. Plus précisément, ce rapport propose l'utilisation de méthodes par apprentissage supervisé et c'est dans cette direction que se sont dirigés les travaux 2019-2020 (Niogret P., 2020 **[3]**).

Le précédent stagiaire a laissé un code fonctionnel, mais non commenté dans le langage R - langage non utilisé à la métropole. Avec le *RMSE*[4] (Root Mean Square Error) comme métrique, un premier algorithme arrivait à prédire une année de pose dans une plage de 18 ans. Pour les matériaux en revanche, aucun élément ne permettant de qualifier la donnée n'a été implémenté. Malgré le fait qu'il n'y a aucune métrique associée à la classification, il ressort que l'algorithme surestime la proportion de collecteurs en béton dans le jeu de données.

Cette méthode d'estimation telle qu'elle est implémentée, mets environ 4 h à fournir des résultats. Comme pour de nombreux modèles de machine learning, il existe des paramètres appelés hyperparamètres afin de calibrer et de régler le modèle. Pour optimiser les résultats, deux hyperparamètres ont été testés séquentiellement afin de trouver le couple ayant les meilleurs résultats.

Le langage utilisé pour la méthode implémentée ne correspond cependant pas aux compétences de l'unité qui va l'utiliser. Le temps de calcul est par ailleurs trop long pour pouvoir travailler correctement dessus. Cela s'explique par le fait que les ordinateurs de la métropole ne sont pas dimensionnés pour effectuer des opérations lourdes. De plus, ne pas pouvoir qualifier un modèle le rend inutilisable. La donnée des matériaux produite ne peut donc pas servir. Le fait de n'avoir fait varier que deux hyperparamètres fait certes gagner du temps de calcul, mais ne permet pas de limiter au mieux le sur-apprentissage d'un modèle et peut expliquer la surreprésentation du béton observé dans la figure 2.

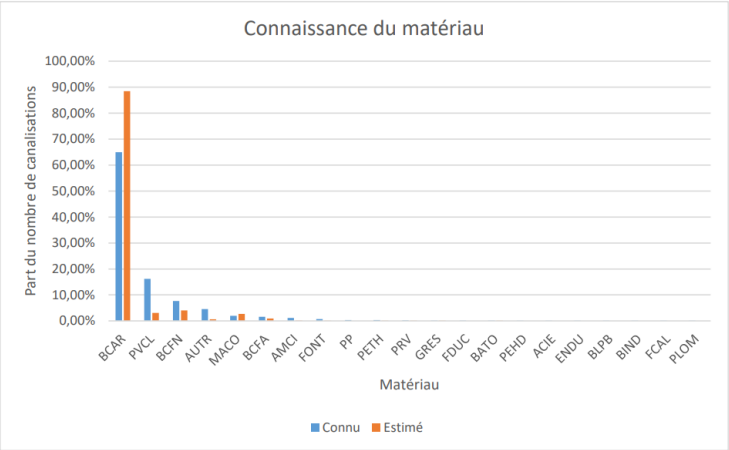



*Figure 2 Estimation des matériaux lors des travaux 2019-2020* **[3]**

Le second travail (Le Moyec T., 2021 **[4]**) ne portait pas entièrement sur l'estimation des années de pose. L'analyse et la reprise des scripts des travaux de 2019-2020 en vue de les améliorer ne représentaient qu'une mission de l'alternant. Ces travaux représentent aussi les premiers tests de redéveloppement des méthodes d'estimations dans un autre langage : Python.

[4] Voir glossaire

Selon les justifications de l'alternant, un changement d'outils s'imposait, et ce, à juste titre. C'est pour cela que les travaux sont passés sur le langage Python. C'est un langage connu de l'équipe et compatible avec Google Colab, outil permettant de lancer du code dans le cloud s'affranchissant ainsi des capacités limitées des ordinateurs à disposition. Ici, seul un script sur les matériaux a été fourni. Celui sur les années de pose est resté inchangé (en R). Il faut cependant préciser que la méthode implémentée pour les matériaux n'est pas une reprise des scripts des travaux 2019-2020. Pour estimer les matériaux, un arbre de décision et une régression logistique ont été implémentés. L'algorithme faisait tourner les deux modèles et renvoyait celui avec les meilleurs résultats à savoir toujours ceux de l'arbre de décision. Comme le regroupement des classes n'avait pas été une idée proposée dans les travaux précédents, il a fallu faire tourner le modèle avec les nouvelles classes. On obtient alors les résultats du tableau 3.

| Arbre de décision | | | | | | | |
|---|---|---|---|---|---|---|---|
| index | Précision | Rappel | F-Score | index | Précision | Rappel | F-Score |
| AMCI | 0,650 | 0,820 | 0,730 | Pondéré | 0,670 | 0,580 | 0,590 |
| AUTR | 0,125 | 0,085 | 0,100 | Macro | 0,596 | 0,583 | 0,585 |
| BTAM | 0,700 | 0,603 | 0,640 | | | | |
| BTAU | 0,805 | 0,790 | 0,795 | | | | |
| FON | 0,565 | 0,485 | 0,520 | | | | |
| PLAS | 0,626 | 0,644 | 0,634 | | | | |
| ROCH | 0,703 | 0,657 | 0,673 | | | | |

*Tableau 3 Résultats de précision, rappel et f-score d'estimation des matériaux des travaux de 2020-2021*

Avec une précision et un rappel de 0.59 et 0.58 cette méthode n'est pas assez viable pour estimer les matériaux de la métropole de Lyon. On remarque cependant que la classe ayant le plus de support à savoir le béton a un F-score 0.79 qui est acceptable.

Un script des matériaux est en python et fourni une première base. L'ancien script en R pour les années de pose donne également un intervalle de 18 ans. Le but est donc d'améliorer ces résultats.

## 2.3. Analyse du problème au regard de l'apprentissage supervisé

Des travaux précédents, il apparaît que tout n'est pas à garder. Le passage de toutes les méthodes en Python et donc à Google Colab est mandatoire. Il faut optimiser les temps de calcul et profiter d'une interface plus propice à être utilisée par les agents de la métropole.

Un état de l'art des méthodes à envisager est à refaire. L'algorithme implémenté en 2019-2020, s'est fait sur les conseils des chercheurs du projet Hireau et est une piste à creuser. Tandis que l'arbre de décision a été implémenté par manque de temps comme solution par défaut.

Le choix de la méthode d'estimation repose sur une théorie mathématique pour qualifier la faisabilité et la difficulté de la résolution du problème à l'aide de l'apprentissage machine. L'utilisation de l'algorithme en tant que tel a été justifiée par des travaux de benchmarking.

### 2.3.1. Présentation de l'apprentissage supervisé

Une des méthodes d'apprentissage automatique est l'apprentissage supervisé. Il consiste à prédire à partir d'informations labellisées des informations non-labellisées à l'aide d'une fonction de prédiction. Dans cette branche de l'apprentissage machine, deux catégories sont à distinguer. Si la variable à prédire est une variable qualitative comme des classes de matériaux, c'est un problème de classification. En revanche, si la variable à prédire est une variable quantitative comme des années de pose, c'est un problème de régression. Il y a une hypothèse à faire sur les données de base ou données d'entraînement. Cette base d'apprentissage, qui est un échantillon de la donnée totale, doit en être représentative. Cela permet à la fonction de prédiction de faire des estimations correctes sur les données qui ne sont pas dans la base d'apprentissage.

La donnée en entrée, appelée ensemble des observations ou ensemble d'entrainement, est un ensemble de $n$ vecteurs de dimension $m$ :

$$X = \{x_1, \dots, x_n\} : \ x_i \, , i \in \{1, \dots, n\} \, , (n, m) \in N^2 , x_i \in \ R^m$$

Auquel est associé l'ensemble des réponses ou labels

$$Y = \{y_1, \dots, y_n\} : \ y_i \, , i \in \{1, \dots, n\} \, , n \ \in N$$

Où $y_i \in \{1, \dots, k\}, k \ \in N$ dans le cas d'une classification multi classes et $y_i \in R$ dans le cas d'une régression.

Le modèle ou classifieur $h : X \to Y$, en apprentissage supervisé, est une structure mathématique où une prédiction $\hat{y}_i$ est faite à partir de l'observation $x_i$. Nous avons donc en entrée pour notre algorithme par apprentissage supervisé un échantillon d'apprentissage $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \in (X \times Y)^n$ et en sortie une règle de prédiction $h : X \to Y \ h \in H, H \subset \ Y^X$. L'objectif est que le modèle h doit être correct sur les données non observées.

Pour définir la qualité de notre hypothèse $h \in H$ on utilise une fonction de perte ou *Loss* qui répond à la question : suivant un exemple $(x, y) \in S$, quelle est la qualité de la prédiction de $h(x) = \ \hat{y}$ ? On définit alors la fonction de perte comme suit :

$$l : H \times (X \times Y) \to \ R^+$$

Il existe cependant autant de fonctions de perte qu'il existe de modèles supervisés en fonction du cas d'utilisation et des données. En voici quelques exemples (Shalev-Shwartz S., 2021 **[5]**) :

$$Perte \ 0 - 1 : l\big(h, (x, y)\big) = \begin{cases} 1 \ si \ h(x) \neq y \\ 0 \ sinon \end{cases}$$

$$Perte \ quadratique : l\big(h, (x, y)\big) = \ (h(x) - y)^2$$

$$Valeur \ absolue \ de \ la \ diff\acute{e}rence : l\big(h, (x, y)\big) = |h(x) - y|$$

Le problème d'apprentissage revient au final à trouver $\min_{h \in H} \mathbb{E}[l(h, (x, y)], (x, y) \in X \times Y$

### 2.3.2. Justification de l'utilisation de l'apprentissage supervisé

Avant d'implémenter des méthodes d'apprentissage supervisé, il faut se demander si cette approche a un sens pour estimer nos données. Des résultats empiriques montrent que l'utilisation de ces méthodes fonctionne. Il est intéressant de répondre à cette question pour justifier de manière théorique l'utilisation de l'apprentissage machine. Il faut être capable de répondre aux questions : de combien d'exemples avons-nous besoin pour apprendre de manière fiable ? Peut-on quantifier la difficulté de l'apprentissage ? Pour étudier la faisabilité de l'estimation des années de pose et des matériaux, il faut se pencher sur un cadre théorique de l'apprentissage automatique : L'apprentissage PAC (*Probably Approximately Correct*) (Kontorovich et al, 2017 **[6]** | J2kun, 2014 **[7]**) et l'apprentissage PAC agnostique (Fort et al, 2020 **[8]**).

Afin de conserver la concision de ces travaux, l'explication et la justification de l'obtention des formules permettant de catégoriser un problème de PAC ou PAC agnostiques est en **annexe A** (p. 72).

Il faut cependant retenir, qu'à partir d'hypothèses de départ différentes pour le modèle PAC ou PAC agnostique (Figure 3), il est possible d'obtenir une valeur caractérisant le nombre minimum d'échantillons à posséder pour que l'algorithme de prédiction soit Probablement Approximativement Correct. Cependant, cette estimation au sens PAC présente plusieurs défauts. Le premier étant qu'on impose une hypothèse extrêmement forte sur l'existence d'une fonction $f$ amenant à une erreur de généralisation nulle. Un moyen de se rendre compte de la contrainte de cette hypothèse est qu'il est impossible d'avoir des données bruitées dans $S$. Sinon f ne peut pas avoir une erreur de généralisation nulle car elle fera forcément des erreurs sur le jeu de test où les $y_i$ ne sont pas connus. Si on prend l'exemple de l'estimation des matériaux, il est possible qu'une canalisation en béton parmi deux identiques voisines ait été remplacée par une autre en plastique. La seule différence entre ces deux canalisations presque identiques est alors le type de matériaux. Il ne peut donc pas exister une fonction de prédiction f déterminant parfaitement le jeu de données. C'est pour cela que le PAC agnostique propose des hypothèses plus faible qui est donc plus difficile à atteindre.

| | PAC | PAC agnostique |
|---|---|---|
| Distribution : | $\mathcal{D}$ sur $\mathcal{X}$ | $\mathcal{D}$ sur $\mathcal{X} \times \mathcal{Y}$ |
| Étiquetage : | $f \in \mathcal{H}$ | pas dans la classe ou n'existe pas |
| Risque | $L_{\mathcal{D},f}(h) = \mathcal{D}(\{x : h(x) \neq f(x)\})$ | $L_{\mathcal{D}}(h) = \mathcal{D}(\{(x,y) : h(x) \neq y\})$ |
| Échantillon $S$ : | $(x_1, \ldots, x_m) \sim \mathcal{D}^m$ $\forall i,\ y_i = f(x_i)$ | $((x_1,y_1), \ldots, (x_m,y_m)) \sim \mathcal{D}^m$ |
| Objectif : | $L_{\mathcal{D},f}(A(S)) \leq \epsilon$ | $L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$ |

*Figure 3 Récapitulatif de la théorie PAC et PAC agnostique. Source :* **[5]**

Ainsi, au niveau des matériaux, nous disposons de 58 245 tronçons aux matériaux connus classés dans 7 classes différentes. Les observations ont chacune 47 attributs[5] pour les caractériser.

[5] Voir section 2.4.1.1 Choix des données en entrée

En prenant une précision $\epsilon$ et un $\delta$ de 0.05, au sens PAC agnostique nous avons donc $\frac{2\log\left(\frac{2|H|}{\delta}\right)}{\varepsilon^2} \approx 33\,057 < 58\,245$. Au niveau des années de pose, nous disposons de 30 139 tronçons aux années de pose connues. Au niveau des classifications, comme H doit être fini, nous prendrons toutes les dates de 1906 date la plus ancienne à 2022, soit 116 classes. Les observations ont chacune 47 attributs pour les caractériser. En prenant une précision $\epsilon$ et un $\delta$ de 0.05 au sens PAC agnostique nous avons donc $\frac{2\log\left(\frac{2|H|}{\delta}\right)}{\varepsilon^2} \approx 78\,905 > 30\,139$. Au sens PAC nous avons $\frac{\log\left(\frac{|H|}{\delta}\right)}{\varepsilon} \approx 1\,966 \ll 30\,139$.

Ainsi, sachant le nombre d'éléments connus et le nombre minimum d'éléments pour qu'un modèle soit apprenable au sens PAC, Nous pouvons valider l'utilisation de méthodes d'apprentissage supervisé pour l'estimation des matériaux. En revanche, pour l'estimation des années de pose, nous pourrons trouver un modèle apprenable au sens PAC, mais pas au sens PAC agnostique. Cela montre que l'apprentissage seul ne permettra pas d'obtenir des résultats probablement approximativement corrects. Cela étant dit, cette conclusion est valide pour un $\epsilon$ et un $\delta$ de 0.05. On pourrait diminuer ces seuils, mais l'objectif était de justifier l'utilisation du machine learning.

### 2.3.3. Choix de l'algorithme d'apprentissage

Au vu des explications précédentes, il est légitime d'utiliser l'apprentissage machine pour déterminer les attributs des collecteurs. Il reste cependant encore à déterminer l'algorithme qui sera utilisé.

Il existe une dizaine de catégories d'algorithmes d'apprentissage supervisé. Cependant, ils n'ont pas tous été modélisés avec le même but. Le choix de l'algorithme se fait notamment à partir des données en entrée. Notre jeu de donnée est composé de données tabulaires. Les labels sont soit des entiers pour les années de pose, soit des chaînes de caractères pour les classes de matériaux. Deux options se posent alors : choisir un algorithme de régression pour les années de pose et un algorithme de classification pour les matériaux. Deux méthodes différentes sont alors développées. Ou trouver un algorithme capable d'effectuer les deux. Il est donc important de trouver un algorithme cohérent avec les problèmes à résoudre.

D'après une étude publiée dans le *Journal of Machine Learning Research* en 2014 (Fernández-Delgado et al, 2014 **[9]**), un ensemble de chercheurs a travaillé sur l'estimation du meilleur algorithme pour traiter des données tabulaires. 179 classifieurs de 17 familles différentes ont été testés sur 121 jeux de données différents. Il ressort de cette étude que l'algorithme le plus efficace était le RandomForest parallélisé. Il était attendu que ce type d'algorithme soit aussi performant. Il appartient en effet aux algorithmes d'apprentissage d'ensemble. Le Random Forest est une combinaison de modèles individuels d'arbres de décisions travaillant sur des jeux de données partiellement différents. Cette approche permet d'améliorer la performance et la stabilité du modèle (maximiser la précision et minimiser la variance) par rapport à chacun des modèles pris séparément. Il est notamment connu pour extraire des informations et des règles de sources de données « obscures », la combinaison d'arbres de décisions permettant de générer des règles métiers invisible sans analyse profonde des données. Cependant, cet algorithme est incapable d'extrapoler des données. Il ne peut donc pas être utilisé dans le cas d'une régression pour estimer les années de pose.

En 2018 (Olson et al, 2018 **[10]**), une autre étude réalisée par d'autres chercheurs a comparé 13 algorithmes sur 165 jeux de données biomédicales : des jeux de données tabulaires. Ils ont également analysé le gain potentiel en paramétrant correctement les différents algorithmes, chose qui n'a pas été faite dans l'étude précédente.

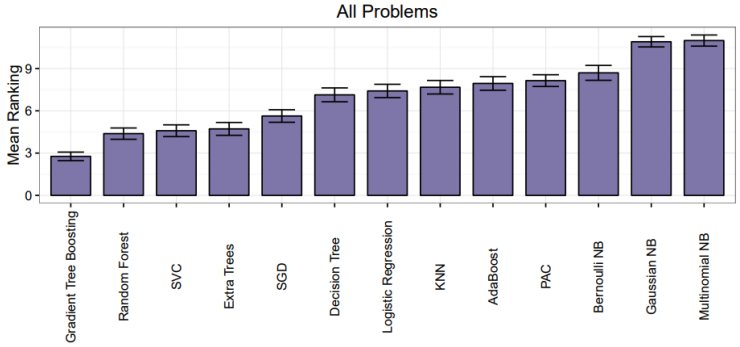



*Figure 4 Rang moyen des algorithmes en termes de performance sur les différents jeux de données* **[10]**

Il ressort de cette étude un algorithme surpassant en termes de rang moyen tous les autres algorithmes : le Gradient Tree Boosting (Figure 4). Il apparaît également comme le modèle surpassant le plus les autres sur l'ensemble des 165 jeux de données. Cet algorithme, est une amélioration de la famille des algorithmes de type arbre de décision.

Ils ont également étudié la qualité du modèle suivant la batterie de paramètres en entrée : *Tuning* d'algorithmes. Ils ont ainsi montré qu'il était possible d'améliorer de manière significative les résultats du Gradient Boosting en trouvant le bon jeu de paramètres.

Cet algorithme de Gradient Boosting a été soumis à des améliorations au fil des années et une de ses améliorations, sortie en 2014 s'appelle le XGboost pour eXtreme Gradient Boosting (XGBoost Documentation, 2014 **[11]** | Hachcham A., 2021 **[12]**). Cette amélioration, rendue célèbre en 2014 a été la plus performante sur un concours de machine Learning organisé par la plateforme Kaggle. Le but du concours était de labelliser correctement des observations du LHC pour aider à la détection du Boson de Higgs parmi toutes les données (Laboratoire de l'Accélérateur Linéaire, 2014 **[13]**).

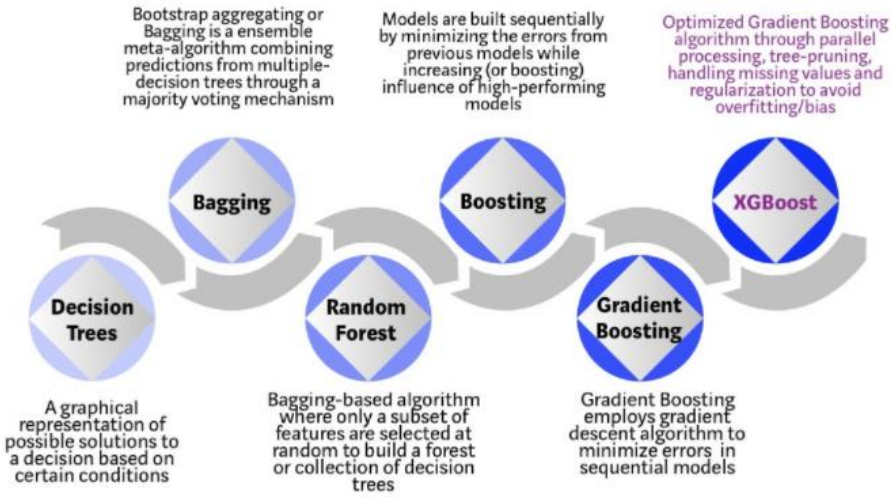



*Figure 5 Évolution de méthodes menant au XGBoost. Source : (Vishal Morde, 2018, XGBoost Algorithm)*

La raison de cette performance est que les méthodes basées sur les arbres traitent les caractéristiques indépendamment les unes des autres et élaborent des règles sur la base des valeurs de ces caractéristiques. En d'autres termes, un algorithme basé sur les arbres tentera de déterminer des attributs basés sur des questions alors qu'un réseau de neurone par exemple essayera des combinaisons du type : ProbaBéton = W1*longueur + W2*profondeur+W3*PositionX. On peut voir qu'avec des colonnes hétérogènes, mélanger linéairement des choses très différentes n'est pas une tâche facile. Cette explication est simplifiée, mais devrait permettre de montrer l'idée. Cela signifie que même s'il n'est pas automatiquement le meilleur choix pour les données tabulaires, XGBoost a du sens en tant que tentative prioritaire, car s'il ne parvient pas à donner de bonnes performances, c'est un indicateur probable que le boosting n'est pas une approche optimale du problème.

Cet algorithme reste en 2022 l'état de l'art dans la famille des arbres de décisions. C'est donc lui qui a été retenu pour estimer les paramètres manquant de la couche des collecteurs.

### 2.3.4. Présentation de l'algorithme XGBoost

Dans les travaux de 2020-2021, un arbre de décision a été implémenté, cet apprenant peut être catégorisé comme « faible » (Stéphanie G., 2021 **[14]**). Un moyen d'améliorer un apprenant faible et d'en associer plusieurs, cela s'appelle l'apprentissage par ensemble. C'est notamment le principe du Random Forest. On apprend à des arbres de décision avec des jeux de données différents à prédire une donnée, puis, on les fait voter pour estimer la donnée finale. Cette méthode de regroupement est appelée Bagging : une méthode ensembliste parallèle.

Une autre approche, celle utilisé par XGBoost est une méthode ensembliste séquentielle : le Boosting. L'algorithme va générer un premier arbre de décision et va attribuer un poids de valeur égale à toutes les observations. Ensuite, si une observation est mal classée, elle verra son poids augmenter. Ensuite et de manière itérative, un modèle n est construit avec en entrée la sortie du modèle n-1. Ainsi, le modèle n va apprendre des erreurs pondérées du modèle n-1. Son rôle est donc de les corriger. Sur des observations $y$, le premier modèle va renvoyer $\hat{y}$ et celui d'après va s'entraîner sur $y - \hat{y}$. Si la prédiction est bonne, $y - \hat{y} \approx 0$ donc le modèle suivant ne modifiera pas énormément les données en entrée. Par contre si la prédiction est mauvaise, $y - \hat{y}$ aura une valeur non-proche de 0 qui aura plus tendance à être modifiée.

L'algorithme XGBoost (Starmer J., 2019 **[15]**) est une optimisation du Gradient Boosting, lui-même dérivé du Boosting, car il utilise un algorithme de descente de gradient pour pondérer les données au lieu de travailler sur $y - \hat{y}$ .

Si l'on reprend les données initiales :

$$X = \{x_1, \dots, x_n\} : \ x_i \, , i \in \{1, \dots, n\} \, , (n, m) \in N^2 , x_i \in \ R^m$$

$$Y = \{y_1, \dots, y_n\} : \ y_i \, , i \in \{1, \dots, n\} \, , n \ \in N$$

Le but de l'entraînement du XGBoost, est de trouver les meilleurs paramètres $\theta$ pour estimer au mieux Y à partir de X. Pour cela il faut définir une fonction objectif dont le but de l'algorithme est de minimiser sa sortie.

$$obj(\theta) = \ L(\theta) + \ \Omega(\theta) \qquad (1)$$

Le terme de régularisation $\Omega(\theta)$ est un élément permettant de limiter l'apprentissage du modèle à partir des données d'entraînement. S'il n'était pas là, on pourrait observer des effets d'*overfitting* : le modèle serait excellent pour prédire des données d'entraînement, mais ne serait pas aussi performant sur de nouvelles données, celles qui nous intéressent. Il permet de réguler l'équilibre entre biais et variance pour avoir un modèle « simple » avec une bonne capacité de prédiction. L'algorithme construit sa prédiction à partir de « weak learners ». Contrairement à un arbre de décision classique, l'algorithme va construire des CART (Classification And Regression Trees) où à chaque feuille sera associé un score permettant de témoigner de l'importance de cette feuille dans la classification.

Pour les années de pose, la fonction de perte utilisée sera le RMSE (Root Mean Square Error), car nous faisons face à un problème de régression. Tandis que pour la classification des matériaux, la fonction de perte utilisé sera la *logloss*. Ces deux fonctions sont des fonctions classiques pour des problèmes de régression et de classification.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{n}}$$

La fonction de perte *logloss* est la log-vraisemblance négative d'un modèle. Plus simplement, elle mesure la vraisemblance. Mais comme elle travaille avec des probabilités, les valeurs deviennent rapidement intraitables pour une machine car trop petites. On utilise alors la fonction log. La négation sert à la rendre positive.

$$y \in \{0,1\}, \quad L_{log}(y,p) = -(y\log(p) + (1-y)\log(1-p)) \quad Logloss\ binaire$$

Dans notre cas avec K labels, on pose $Y \in M_{i,k}(\mathbb{R})_{1\leq i\leq N, 1\leq k\leq K}$ et $P \in M_{i,k}(\mathbb{R})_{1\leq i\leq N, 1\leq k\leq K}$, $p_{i,k} = \Pr(y_{i,k} = 1)$

$$L_{log}(Y,P) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log p_{i,k}$$

Afin de conserver la concision de ces travaux, la construction de la fonction objectif (1) est explicité en **annexe B** (p. 73).

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \qquad (2)$$

Cette équation (2) est donc la fonction objectif qui permet de mesurer la qualité de la structure d'un arbre $q(x)$. Elle dépend du nombre d'arbres T en jeu, ainsi que de la somme des gradients et des hessiennes de chaque arbre (G et H).

Dans l'idéal. Il faudrait alors tester toutes les configurations possibles de CARTs et prendre l'arbre avec la plus petite valeur de $obj^*$. Cette approche n'étant pas réalisable dans un temps fini, l'algorithme propose de construire et d'optimiser un niveau d'arbre à la fois. De manière plus concrète, une feuille contenant assez d'éléments va être divisée en deux feuilles qui seront soumises à une étude de gain.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

Le Gain total de la séparation d'une feuille est donc la somme du gain de la feuille de gauche (Left) et de la feuille de droite (Right) moins le gain de la feuille originelle. Le tout auquel est soustrait $\gamma$ appelé le paramètre de complexité d'arbre. Si le Gain est positif, la séparation de la feuille mère permet de mieux analyser le jeu de données. Si le Gain est négatif, on va élaguer ces deux nouvelles feuilles. Cette étape s'appelle le *Pruning.*

Dans le cas de la régression, $\frac{G_L^2}{H_L+\lambda} = \frac{(Somme\ des\ résidus)^2}{Nombre\ de\ résidus+\lambda}$ ; $\lambda$ est un terme de régularisation. Il a pour but de minimiser la sensibilité de la prédiction aux observations individuelles ou d'entraînement. Le gain d'une feuille est ainsi inversement proportionnel au nombre de résidus dans la feuille. Dès lors, même si $\gamma = 0$, avec un $\lambda = 1$ , il est possible d'obtenir un Gain négatif et donc d'élaguer des branches. Une fois que l'arbre a atteint sa profondeur maximale ou bien que $\lambda$ et $\gamma$ ne permettent plus la création de nouvelles branches, cet arbre est ajouté à la prédiction finale avec un coefficient multiplicateur appelé learning rate $\varepsilon$. Comme son nom l'indique, il permet de modérer l'apprentissage et de le limiter afin d'augmenter la qualité du modèle.

Dans le cas d'une classification, qui ne sera pas détaillé ici, la seule différence est la fonction de perte. L'algorithme travaille ensuite avec des probabilités d'appartenance à chaque classe. Les formules de Gain ou de similarité et la sortie de l'algorithme sont différentes, mais les raisonnements restent identiques.

Dans cette partie a été vu la manière de construction d'un arbre de manière théorique avec la construction complète de la fonction d'optimisation.

L'algorithme possède de nombreux avantages quand l'on travaille sur de gros volumes de données. La combinaison séquentielle de « *weaks learners* » ne permet pas comme un Random Forest de faire apprendre des modèles en parallèle. C'est pour cela que cet algorithme implémente des méthodes intéressantes pour rester performant. Comme vu dans la partie théorique et structurelle de cet algorithme, il est également quasiment immédiat de travailler en classifieur ou en régresseur. Cela permettra de gagner du temps lors de la phase d'implémentation et cela permettra de se focaliser sur d'autres aspects du machine learning.

Comme vu précédemment, il existe des variables ($\varepsilon$, $\lambda$ et $\gamma$ notamment) appelées hyperparamètres qui servent à ajuster le modèle dans ses prédictions. Comme montré dans l'étude de 2018 **[10]**, il apparaît faisable d'améliorer les prédictions en changeant ces paramètres. Comme ce sont des entrées du XGBoost, il est possible d'utiliser des méthodes externes afin de les estimer au mieux. Il existe au total 10 hyperparamètres qui permettent de limiter le sur-apprentissage et d'améliorer les métriques.

La librairie Python qui implémente les briques élémentaires du XGBoost a été largement soutenue par de nombreux contributeurs. De nombreuses fonctionnalités ont donc été implémentées. Notamment la possibilité d'entraîner le modèle sur un processeur graphique. Comme le code tournera sur un Google Colab, il y a donc lieu d'utiliser le GPU à disposition. Cela permettra d'accélerer les temps de calculs.

Il y a un élément qui n'a pas été mentionné dans la présentation de l'algorithme : Le choix des conditions de séparation des feuilles. Quand le jeu de donnée est considéré comme faible, XGBoost utilise un algorithme glouton et teste toutes les possibilités pour maximiser le gain des feuilles. Cependant, si l'on commence à augmenter le nombre d'attributs à étudier et le nombre d'observations, ce type de méthode n'est plus envisageable, car il est trop gourmand en temps et en ressources de tester toutes les possibilités de seuils. Pour palier à ce problème, il est utilisé un algorithme glouton d'approximation. Il travaille en séparant les données de manière intelligente en quantiles pondérés.

### 2.3.5. Importance des données en entrée de l'apprentissage

L'apprentissage machine doit permettre d'atteindre 100 % de complétude dans les années de pose et des matériaux. Cependant, les résultats fournis resteront « la meilleure estimation au vu des données fournies ». Il est possible qu'il existe un biais dans les données ou que le jeu de données ne soit pas assez homogène pour permettre la différenciation correcte de toutes les classes. Ces résultats restent une estimation et pour l'améliorer, il faut une meilleure connaissance réelle du terrain. C'est-à-dire augmenter la quantité et vérifier la qualité des informations en entrée.

Le meilleur moyen pour qualifier les données reste les données d'archives et les connaissances techniques des équipes sur le terrain.

## 2.4. Mise en œuvre pour l'estimation des années de pose et des matériaux

### 2.4.1. Améliorations générales du modèle

#### 2.4.1.1. Choix des données en entrée

L'ajout de données cohérentes avec les attributs à prédire est primordial. Comme vu dans les limites de l'apprentissage machine, tout dépend des données en entrée. Fournir à l'algorithme des données qui n'ont aucune influence sur l'aide à la décision des matériaux ou années de pose peut dans le meilleur des cas ne rien faire et dans le pire, dégrader les prédictions (Baudoux et al, 2021 **[16]**).

Une des données les plus importantes à ajouter est celle de la spatialisation. Dans un contexte métropolitain où le centre de Lyon est beaucoup plus ancien que toutes les communes en périphérie, l'information de la position des collecteurs permettra d'identifier des quartiers anciens ou historiques ou des zones industrielles récentes. Cet élément n'a pas été mis en place dans les travaux précédents et le faire augmentera les performances. Il a donc été ajouté les positions X et Y des centroïdes et des positions amonts et avals des tronçons des collecteurs.

En travaillant avec les équipes, sont apparues des règles métiers qui peuvent être transcrites avec des attributs. Les collecteurs en fonte sont normalement ceux situés sous les ponts et le long des cours d'eau. Cette information peut être remontée aux collecteurs sous forme de distance à l'eau en se servant de la couche des éléments aquatiques de l'un des thèmes de la base de données. En travaillant avec les données d'autres couches SIG à disposition, il a été ajouté la profondeur d'enfouissement de chaque collecteur. Cet élément est important, car il caractérise la résistance du matériau qui doit être utilisé. Enfin, afin d'avoir une vision plus macroscopique que la position des collecteurs exacte, le code INSEE des communes et des arrondissements a été ajouté. Il permettra de rendre compte de politiques d'urbanisations différentes et ou de forcer certains types d'attributs à des zones étendues spécifiques. Notamment, aider à localiser le vieux centre de Lyon afin d'estimer des matériaux anciens et des années de pose anciennes.

Au niveau des variables catégorielles, elles ont subi un encodage *One-Hot*. I.e. une variable à n états possible est transformée en n variables à 1 état possible et ce pour correspondre au modèle d'entré de l'algorithme. Au final, il y a 47 variables décrivant les collecteurs de la Métropole de Lyon.

En faisant tourner l'algorithme XGBoost, il est possible de faire ressortir l'importance des attributs dans l'aide à la décision des matériaux et des années de pose. Il apparaît clairement que la position géographique des tronçons est primordiale dans l'estimation des paramètres.

Au niveau du processus d'estimation des années de pose, les estimations des matériaux viennent ajouter de l'information supplémentaire. Comme les années de pose un attribut moins peuplé que celui des matériaux, les choses ont était faites dans ce sens pour espérer améliorer les prédictions.

#### 2.4.1.2. Cross Validation

Une méthode commune pour qualifier l'estimation de données par apprentissage machine est le *Train-Test-Val*. Les données sont découpées en deux : les données labellisées (matériaux ou années de pose) et les données non labellisées (VAL). Pour éviter le sur-apprentissage, i.e. apprendre parfaitement le jeu labellisé et mal prédire le jeu non labellisé, le jeu labellisé est redécoupé en deux parties. Le plus souvent 20 % / 80 % où 80 % du jeu servira à entraîner le modèle et 20 % à le tester puisque l'on possède à la fois la prédiction et la vérité labellisée. C'est d'ici que viennent les métriques des modèles. Cependant, si les données disponibles sont limitées et hétérogènes, lors du découpage, il peut y avoir des informations ou des attributs qui n'apparaissent pas dans l'un des trois jeux. De plus, selon le découpage 80/20, les jeux peuvent posséder un biais.

Pour pallier cela, il faut utiliser la validation croisée en K-plis (Shahul ES., 2021 **[17]**) (Figure 6). Le jeu est découpé en K plis. K-1 plis serviront à entrainer le modèle et le dernier à le tester. Un modèle sera entraîné sur chaque configuration K-1/K et il en résultera des métriques sur chaque pli du jeu de données. Pour les matériaux et les années de pose, afin d'équilibrer le rapport gain/temps de calculs, K = 7.

Il reste ensuite à agréger toutes les métriques des K-pli pour avoir une vision globale de la performance du modèle sur l'ensemble du jeu de données. Cette méthode bien que plus longue permet de minimiser le biais de l'algorithme et assure les meilleurs résultats possibles sur le jeu de validation.

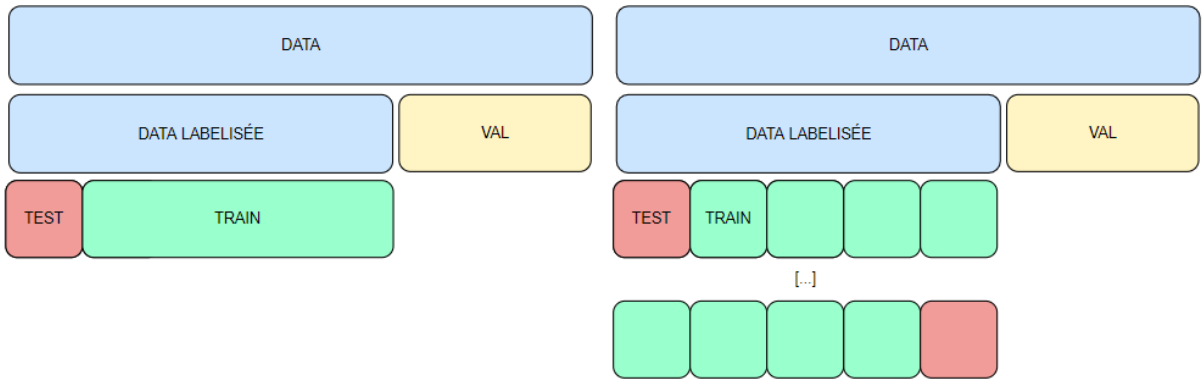



*Figure 6 Différence d'approche de l'apprentissage (Train-Test VS 5-plis)*

### 2.4.1.3. Optimisation hyperparamètres

Comme vu précédemment, l'algorithme possède des hyperparamètres : variables servant à paramétrer l'apprentissage, notamment à réguler le risque de sur-apprentissage. Ils sont au nombre de 10 et l'étude (Olson et al, 2018 **[10]**) a permis de montrer que leur optimisation a un intérêt dans la recherche de performances. Ils sont synthétisés dans la figure 7 ci-dessous.

| Name | Default | Range | Effect | Notes/Tips |
|---|---|---|---|---|
| **n_estimators** | 100 | [1, inf) | Increasing may improve scores with large data. | The number of trees in the ensemble. |
| **learning_rate alias:eta** | 0.3 | [0, inf) | Decreasing prevents overfitting. | Shrinks the tree weights in each round of boosting. |
| **max_depth** | 6 | [0, inf) | Decreasing prevents overfitting. | The depth of the tree. 0 is an option in a loss-guided growing policy. |
| **gamma alias: min_split_loss** | 0 | [0, inf) | Increasing prevents overfitting. | Low values, usually lower than 10, are standard. |
| **min_child_weight** | 1 | [0, inf) | Increasing prevents overfitting. | The minimum sum of weights required for a node to split. |
| **subsample** | 1 | (0, 1] | Decreasing prevents overfitting. | Limits the percentage of training rows for each boosting round. |
| **colsample_bytree** | 1 | (0, 1] | Decreasing prevents overfitting. | Limits the percentage of training columns for each boosting round. |
| **lambda** | 1 | [0, inf) | Increasing prevents overfitting. | L2 regularization of weights. |
| **alpha** | 0 | [0, inf) | Increasing prevents overfitting. | L1 regularization of weights. |
| **missing** | None | (-inf, inf) | Finds optimal null values. | Replace null values with numerical value like -999.0, then set equal to -999.0. See Chapter 5, XGBoost Unveiled. |

*Figure 7 Tableau des hyperparamètres de l'algorithme XGBoost. Source :* **[11]**

Les travaux de 2019-2020 ont ébauché une optimisation des hyperparamètres. Le *learning_rate* pondérant l'ajout de nouvel arbre et le *max_depth* paramétrant la profondeur maximale (nombre de question) de chaque arbre avaient chacun 4 valeurs possibles. Le reste des hyperparamètres ayant gardé leurs valeurs par défaut. L'algorithme bouclait ensuite sur toutes les possibilités et ressortait la meilleure. Cette optimisation réalisée sur les années de pose uniquement a permis d'améliorer le RMSE de 4 %.

Afin de tirer au mieux parti de tous les hyperparamètres pour améliorer la capacité de prédiction du modèle, l'état de l'art en termes d'optimisation d'hyperparamètres a été utilisé (Optuna Documentation, 2018 **[18]**) (**annexe F** p.80). Cette méthode (Figure 8) se base sur l'élagage d'itérations d'entraînement jugé infructueux. Il faut premièrement constituer un univers des hyperparamètres, i.e. donner les plages et catégories de chaque hyperparamètre. On fait ensuite tourner plusieurs modèles en parallèle avec des batteries de paramètres différents. À chaque itération, soit à chaque rajout d'arbre, une fonction objectif est calculée (loss). Si le nouvel essai a un score supérieur à la médiane des X essais précédents à la même itération, cet essai est avorté. Cela permet de gagner beaucoup de temps de calcul et de tester un univers des paramètres très étendu. Ensuite, les X meilleurs essais aident à déterminer par descente de gradient, dans quelle direction diriger chaque hyperparamètre afin d'optimiser la fonction objectif de manière itérative. D'autres méthodes existent mais celle-ci a été retenue pour son temps d'exécution et ses performances (Imamura H., 2020 **[19]**).

Toutes ces étapes sont réalisées avec des jeux d'entraînement et de tests aléatoires afin de ne pas subir de sur-apprentissage. La batterie finale de paramètres permet donc d'étudier l'entièreté du jeu de données de manière optimale. Cette méthode a permis sur les années de pose et des matériaux d'améliorer les métriques de 9 % environ soit un résultat cohérent avec l'étude faite en 2018 **[10].**

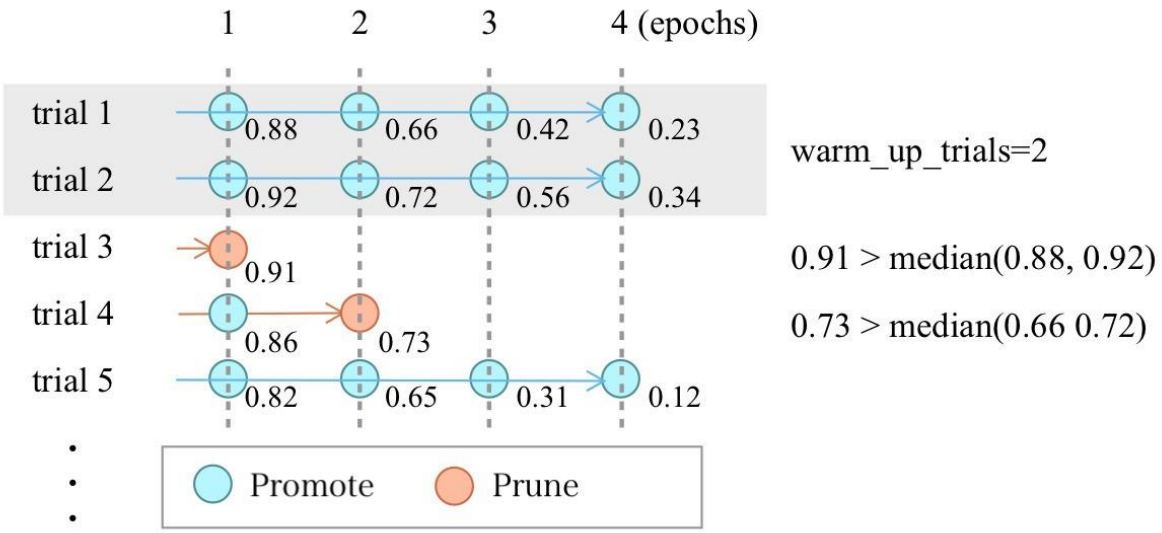



*Figure 8 Présentation de l'élagage des itérations. Source : (Masashi SHIBATA,2021 Optuna)*

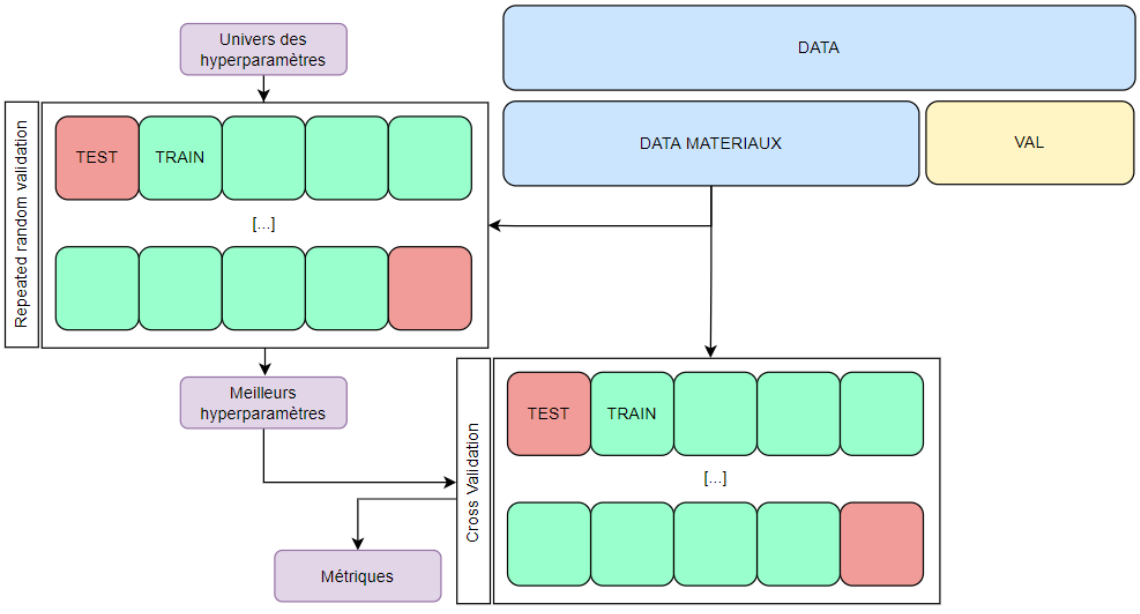



*Figure 9 Processus d'optimisation des hyperparamètres et récupération des métriques*

### 2.4.2. Mise en œuvre pour la reconstitution des années de pose

#### 2.4.2.1. Fonctionnement

Une fois les bons hyperparamètres utilisés et que le modèle a été construit et validé, il reste à estimer la valeur des données qui n'ont pas été labellisées. Pour cela, il faut soit entraîner un modèle en Train-Test-Val sachant les métriques qu'il va donner validées par validation croisée, soit prendre un des modèles déjà entraîné lors de la validation croisée.

En utilisant les arbres construits, par entrée (observation), la sortie pour une estimation par régression correspond à la feuille finale sur laquelle l'entrée s'arrête. La valeur de sortie dépend alors de la somme des résidus ainsi que leur nombre sur la feuille finale :

$$Valeur\ de\ sortie\ = \frac{Somme\ des\ résidus}{Nombre\ de\ résidus + \lambda}$$

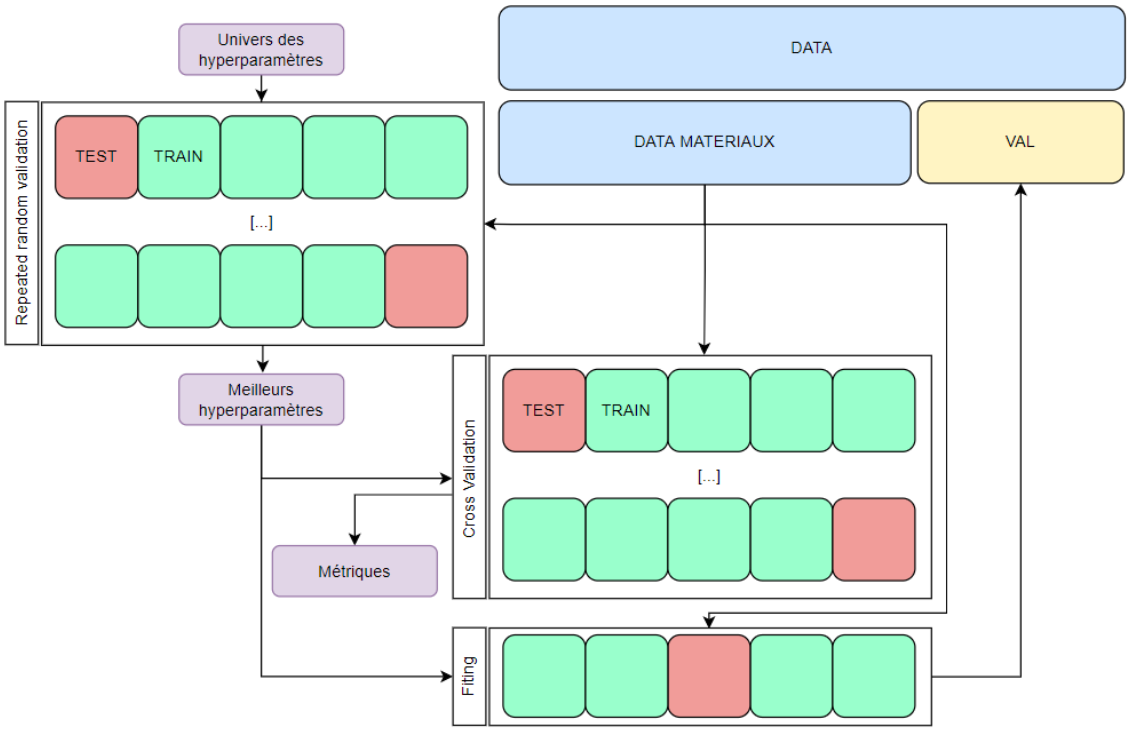



*Figure 10 Processus d'estimation avec ré entrainement d'un modèle*

#### 2.4.2.2. Résultats et discussion

L'estimation des années de pose par apprentissage supervisé permet de passer d'une connaissance de 30,4 % à 100 %. Cependant, les données estimées sont fournies avec une qualification : le RMSE. Cette métrique permet de donner un intervalle de confiance à la donnée estimée. Avec tous les moyens présentés et mis en œuvre il est de 5.8. Une date estimée par machine learning a donc une précision de ± 5,8 ans.

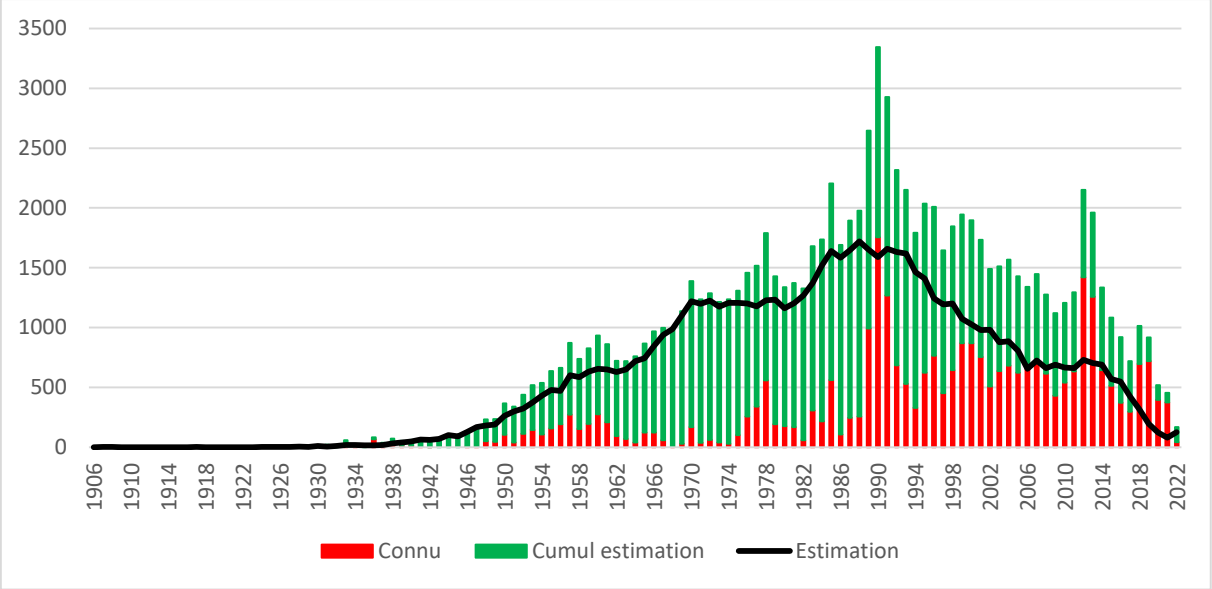



*Figure 11 Histogramme des données connues et estimées*

En étudiant l'histogramme des dates des tronçons estimés et connus (Figure 11), il apparaît que la majorité des données connues se situe après les années 1990. C'est aussi dans ces périodes que sont apparus les premiers systèmes d'informations géographiques avec la rentrée en base de données de plus d'informations. Le manque de connaissances dans les années antérieures à 1960 est normal, les collecteurs placés à cette époque ont plus de soixante ans, âge auquel il faut normalement les remplacer. De plus, c'est en 1958 que l'article L 133 du code de la santé publique oblige les collectivités à se munir d'un réseau d'assainissement. Cette loi peut permettre d'expliquer la forte croissance des collecteurs estimés posés entre les années 1958 et 1970. On remarque également que les prédictions correspondent bien à l'expansion des réseaux urbains Lyonnais entre les années 70 et 90. Les collecteurs estimés après 1990 sont nombreux et peuvent correspondre avec la directive ERU-91 qui oblige les collectivités à collecter et traiter les eaux usées. Après les années 2014, le nombre de collecteurs estimés chute rapidement et c'est un effet cohérent avec le fait que la date des nouveaux collecteurs est quasi-systématiquement renseignée.

Il faut aussi noter que cet histogramme présente une chose très importante. Les données estimées ne suivent pas la distribution des données connues. La modélisation semble pertinente, car on peut lui rattacher des éléments extérieurs au réseau comme des lois ou des tendances au développement qui n'étaient pas visible dans les données connues.

Pour valider les résultats de manière encore plus concrète, il reste à comparer les données à des contrôles ou à des vérités terrain (Figure 12). En 1995, la loi 95-101 ou loi Barnier oblige les collectivités à fournir un rapport annuel sur l'état, le fonctionnement, les missions et les performances du service d'eau potable et d'assainissement. Grâce à ces rapports archivés, il est possible de récupérer le nombre de kilomètres de réseau chaque année et de construire un cumul du linéaire du réseau d'assainissement de la métropole de Lyon. De plus, Gilles Chuzeville a travaillé aux archives et a remonté des cumuls linéaires pour des périodes postérieures à 1995. Grâce à ces éléments et en réalisant les cumuls par années de données connues et estimées, il est possible de croiser les deux informations. Il faut cependant faire attention à l'ajout de territoires gérés en assainissement au cours de l'histoire.

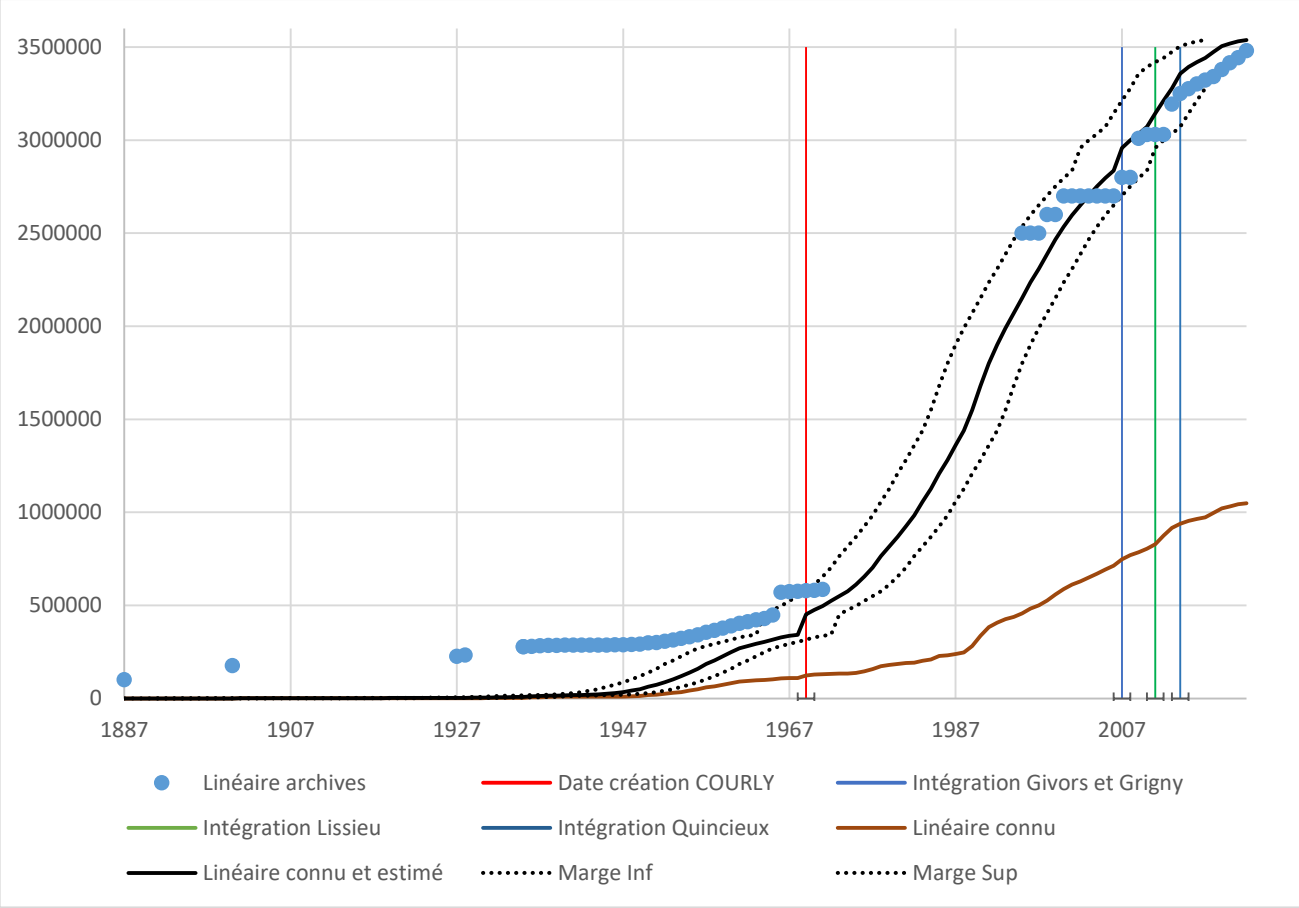



*Figure 12 Comparaison cumul linéaire d'archives et linéaire par année connu et estimé*

La courbe des points du rapport Barnier (post 1995) suit bien la courbe du cumul linéaire des années connues et estimées et reste dans la marge d'erreur de ± 5,8 ans. Cependant, avant la création en 1969 de la communauté urbaine de Lyon (Courly), les données semblent décalées par rapport aux données d'archives. Cela peut s'expliquer par le fait que ces tuyaux ont été remplacés et que peu de tuyaux datant d'entre 1927 et 1968 soient encore en état de fonctionnement aujourd'hui.

Ces résultats ont été obtenus en rajoutant des données en entrée comme vu dans la partie dédiée, mais il a aussi été rajouté les matériaux estimés par apprentissage machine. Sans ces matériaux, le RMSE monte à 6,8 ans. L'estimation des années de pose a été présentée avant celle des matériaux car plus simple à appréhender. Cependant, dans l'ordre chronologique, les matériaux ont d'abord été estimés puis ont servi à estimer les années de pose. Comme l'on connaît davantage de matériaux que d'années de pose, il était plus intéressant de faire les choses dans cet ordre.

La figure 13 page suivante présente les années de poses connues de la Métropole de Lyon. Les années ont été regroupées par tranches de 20 ans afin de faciliter le repérage de zones urbaines avec des canalisations anciennes. La figure 14 elle, représente les données connues avec les données estimées par l'algorithme. On observe correctement la délimitation de Lyon grâce aux années ainsi que la construction de tout le réseau récent au niveau de la confluence du Rhône et de la Saône.

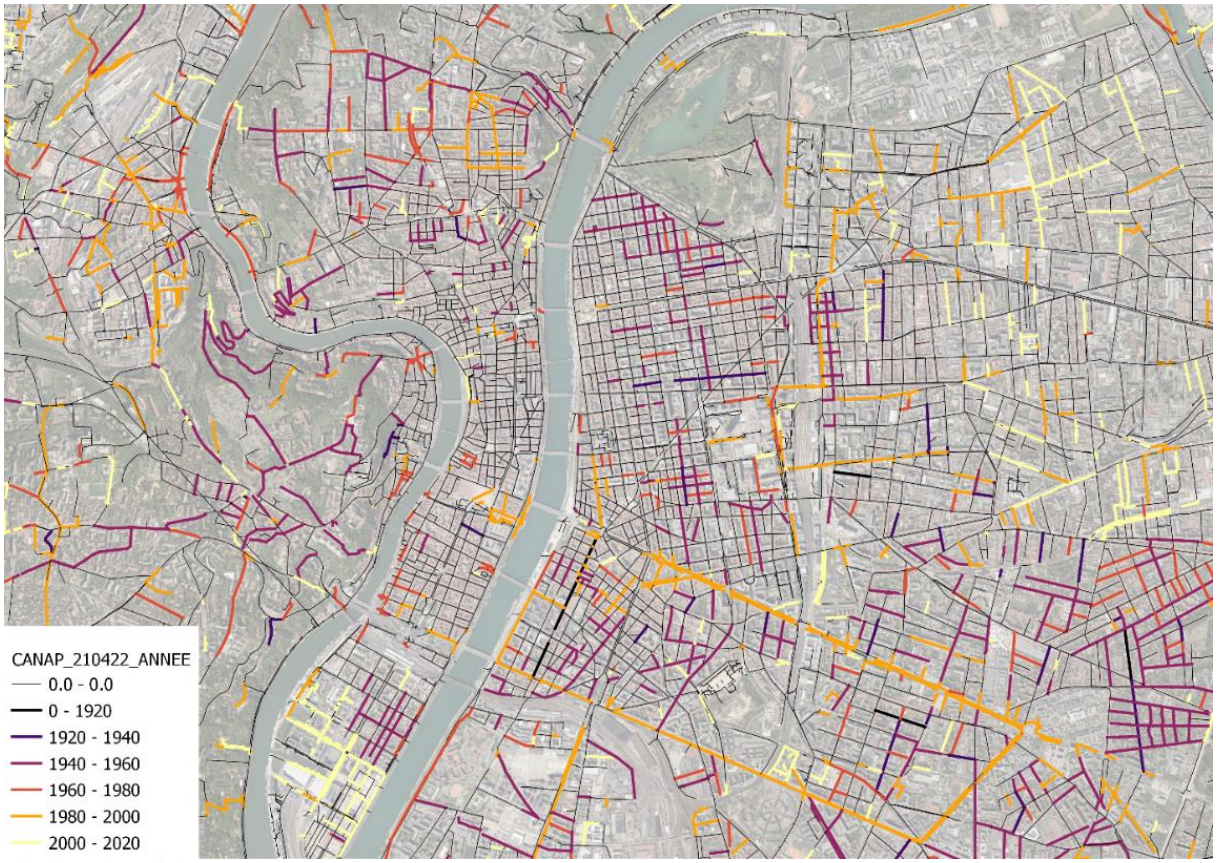



*Figure 13 Export des années connues au 21/04/2022*

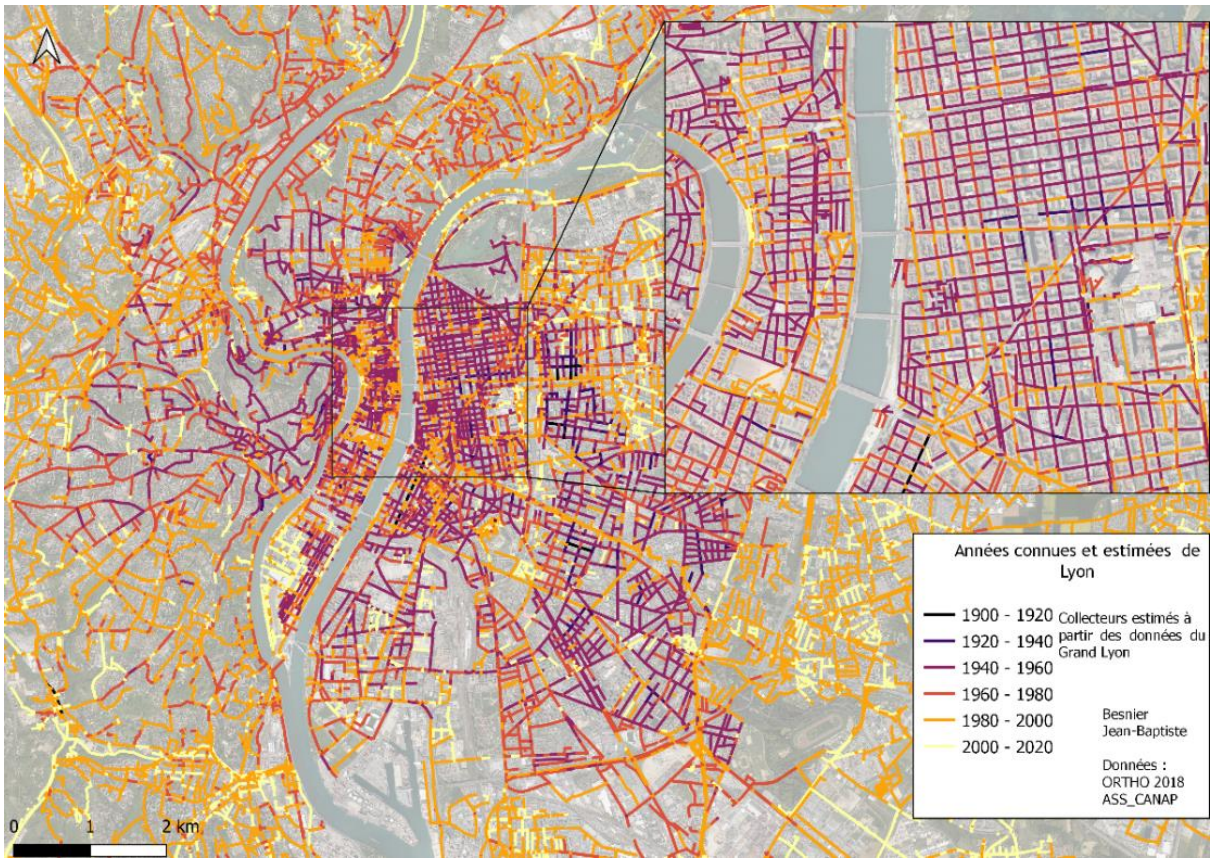



*Figure 14 Estimation des années de pose par apprentissage machine. RMSE : 5.8*

### 2.4.3. Mise en œuvre pour la reconstitution des matériaux des collecteurs

#### 2.4.3.1. Fonctionnement

L'algorithme de prédiction Xgboost fonctionne de la même manière que présenté précédemment. Les améliorations générales restent valides pour estimer les matériaux. Cependant, il faut maintenant résoudre un problème de classification. La sortie de l'algorithme est un vecteur $P \subset [0,1]^7$ car il y a sept classes de matériaux à prédire. Chaque élément i du vecteur donne la probabilité de l'observation d'appartenir à la classe i. Il suffit ensuite de récupérer la classe avec la plus forte probabilité pour estimer le matériau.

Cependant, contrairement aux années de pose où l'on peut prédire des années qui ne sont pas présentes dans le jeu de départ puisque c'est une régression, ici les améliorations mises en place ne suffisent pas à régler le problème de déséquilibre des classes. En reprenant le tableau de regroupement des données du tableau 2, une classe est prépondérante dans le jeu de données. La classe des bétons avec métaux BTAM représente 61 % des matériaux. L'algorithme aura donc en entrée une surreprésentation d'une classe et pourra avoir un biais en la surreprésentant dans les résultats. Des classes avec des supports plus faibles comme celle regroupant les fontes (1 %) ou celle de l'amiante (1 %) pourrait simplement disparaître des prédictions.

Pour valider le modèle, la précision et le rappel seront utilisés, métriques usuelles dans des problèmes de classification.

| Classe | Précision | Rappel | F-Score | Moyenne | Précision | Rappel | F-Score |
|---|---|---|---|---|---|---|---|
| AMCI | 0,650 | 0,820 | 0,730 | Pondéré | 0,670 | 0,580 | 0,590 |
| AUTR | 0,125 | 0,085 | 0,100 | Macro | 0,596 | 0,583 | 0,585 |
| BTAM | 0,700 | 0,603 | 0,640 | | | | |
| BTAU | 0,805 | 0,790 | 0,795 | | | | |
| FON | 0,565 | 0,485 | 0,520 | | | | |
| PLAS | 0,626 | 0,644 | 0,634 | | | | |
| ROCH | 0,703 | 0,657 | 0,673 | | | | |

*Tableau 4 Résultats de précision, rappel et f-score d'estimation des matériaux des travaux de 2020-2021*

Comme vu dans le tableau 4 ci-dessus synthétisant les résultats des travaux 2020-2021, la précision des classes les plus présentes sont les plus élevées. Il faut cependant noter que l'amiante est bien caractérisé par l'ancien modèle.

### 2.4.3.2. Améliorations nécessaires aux matériaux

#### 2.4.3.2.1. Pondération

La première amélioration pour limiter le déséquilibre des classes est de pondérer les classes afin de leur donner plus d'importance (Wang et al, 2021 **[20]**). Il est souhaitable par exemple d'associer un poids élevé à la classe FON et un poids faible à la classe BTAM et que si l'algorithme hésite entre les deux, le poids lié à FON fasse en sorte de garder FON à la fin.

Pour cela, pour chaque classe, un poids inversement proportionnel au nombre d'élément la constituant a été calculé. Il est ensuite associé à chaque observation labellisée et suit les observations tout au long de l'entraînement. Il est ensuite utilisé pour pondérer les hessiennes et les gradients dans la fonction à optimiser. Il a donc exactement le rôle souhaité : si un poids fort est associé à une observation, la fonction objectif sera meilleure et inversement pour un poids faible, donnant plus d'importance aux classes sous représentées.

Cette amélioration couplée à celles commune aux années de pose a permis d'augmenter le F-score de 58,5 % à 83,5 % par rapports aux travaux de 2020-2021

#### 2.4.3.2.2. Synthétisation et suppressions des données

Pour limiter davantage la différence d'éléments entre chaque classe, une méthode tirée de l'apprentissage dans le domaine de la finance a été utilisée. Il existe de nombreux cas où le déséquilibre des classes est plus élevé que dans cette situation des matériaux. Détecter une fraude financière par exemple, requiert de s'entraîner sur un jeu déséquilibré à hauteur de 99 % Non-fraude / 1 % fraude. Pour pallier ce problème, la synthétisation de données est utilisée. Les classes les moins représentées se verront ajouter des éléments artificiels créés à partir des données de départ grâce à l'algorithme SMOTE (*Synthetic Minority Over-sampling Technique*) (Chawla et al, 2002 **[21]**). De façon brève, cet algorithme sélectionne un point de classe A sur le plan, cherche ces voisins, en sélectionne un au hasard et vient créer un point synthétique à une distance aléatoire entre ces deux points. Dans le cas de notre classification, c'est la même chose, à la dimension du problème prêt.

Cette amélioration couplée à celles communes aux années de pose et à la pondération a permis d'augmenter le F-score de 58,5 % à 84,0 % par rapport aux travaux de 2020-2021

Il a également été implémenté l'algorithme SMOTEEN qui a pour rôle de supprimer des éléments trop ambigu pour limiter le recouvrement de classes entre elles. Mais les résultats, bien que meilleurs que ceux de bases, sont inférieurs au modèle sans suppression de données.

Il faut cependant rester vigilant lorsque l'on synthétise de la donnée. Si l'on procède à un ré-échantillonnage et qu'on valide le modèle par validation croisée K-plis, certains auront des échantillons en commun et le modèle sera surestimé **[17]**. Il faut premièrement diviser les données en plis correspondants, puis procéder aux surs échantillonnages à l'intérieur de la boucle d'entraînement et de test. Cependant, le ré échantillonnage doit être fait uniquement sur les données d'entraînement. Les données de test doivent, elles, rester le plus fidèle possible au jeu de validation.

#### 2.4.3.2.3. CleanLab

La détermination de matériaux post-travaux est réalisée par les équipes inspections télévisées ainsi que par les géomètres et les équipes d'exploitation réseau. Il est cependant à noter qu'il n'est pas toujours aisé de déterminer les matériaux d'une canalisation à cause de son état, de sa propreté ou d'un enduit recouvrant l'intérieur. Cependant, les informations des matériaux de la base de données sont tirées de devis de travaux et d'inspections post-travaux. Néanmoins, des collecteurs de la base de données peuvent-être mal labellisés. Pour obtenir les meilleurs résultats en apprentissage supervisé, la cohérence et la propreté des données sont des éléments très important. Réussir à trouver les collecteurs mal labellisés et à les retirer de l'entrainement pourrait permettre de mieux distinguer les classes de matériaux.

Pour cela, en février 2022, une start-up utilisée par notamment Google, Amazon et Facebook nommée CleanLab a mis à disposition de tous, ses outils et propose de donner à chaque observation un score de qualité afin de trier les observations ayant potentiellement des problèmes de labellisation (Northcutt C, 2022 **[22]**). Les labels les plus probables sont alors suggérés.

Le fonctionnement (Figure 15) ne sera pas explicité ici, le lecteur est invité à aller lire la publication d'avril 2021 (Northcutt et al, 2021 **[23]**). L'algorithme commence par s'entraîner par validation croisée sur le jeu de données. Des matrices de probabilités sont construites pour ensuite sélectionner les observations ayant une probabilité d'appartenance à une classe supérieure à un seuil dépendant du nombre d'éléments de la classe estimé par le modèle et de la somme des probabilités des éléments prédit de cette classe.

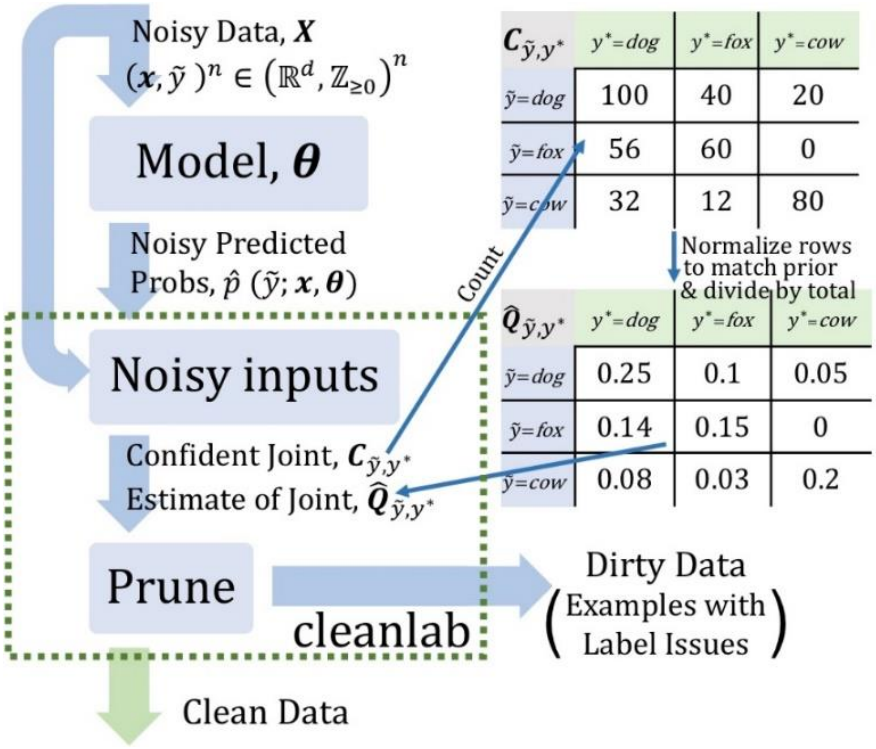



*Figure 15 Principe de fonctionnement CleanLab.* **[23]**

Lors de la sortie publique des outils en février 2022, seule une bibliothèque python était publiée, depuis, une plateforme d'analyse en ligne a été mise en place, mais n'a pas été utilisé pour ces travaux. CleanLab a été imaginé dans l'idée de travailler avec des données de plusieurs millions d'observations. La conclusion de la publication de 2021 statuait ses résultats sur des données images avec des réseaux de neurones et conclut l'utilité de s'intéresser à d'autres modèles comme le XGBoost par exemple. Cependant, avec un peu moins de 100 000 tronçons, le jeu de données est trop faible pour que les probabilités jointes soient fiables. De plus, le problème de déséquilibre des classes pose ici encore plus de problèmes. Ainsi, CleanLab va enlever beaucoup trop d'éléments qui ne sont pas réellement des erreurs de labels et proposer la classe majoritaire dans la plupart des cas.

Cette méthode permet cependant de remonter la probabilité de jointure intra-classe. C'est-à-dire que l'on peut obtenir une mesure du recouvrement entre deux classes. Et ainsi justifier le regroupement des classes qui a été fait lors de la préparation des données pour passer à sept classes de matériaux.

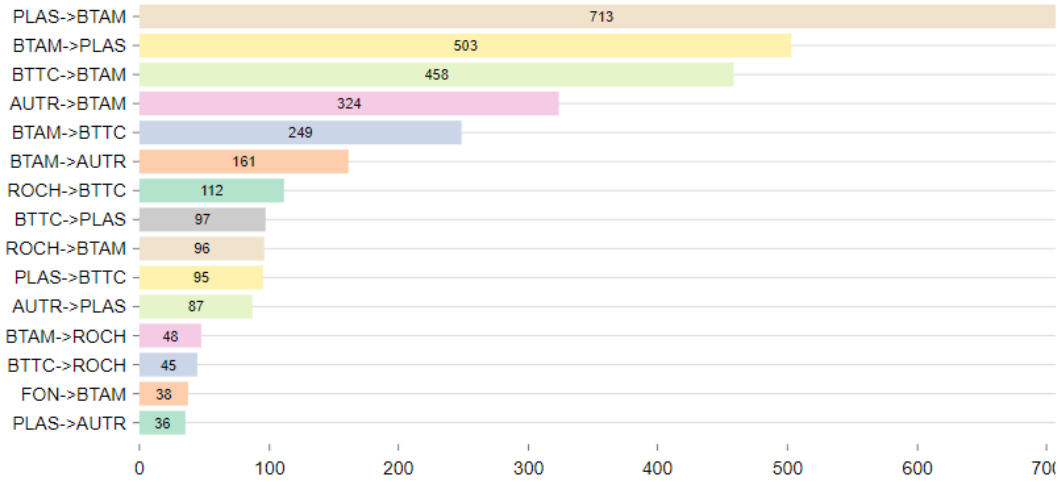



*Figure 16 Histogramme des similarités entre classes*

Le graphique (Figure 16) ci-dessus présente les éléments d'une classe A ayant de fortes similarités avec des éléments de la classe B. On peut observer que les classes avec le plus d'éléments en commun sont les classes de béton avec métaux et les plastiques, classes les plus présentes du jeu de donnée. C'est également compréhensible au niveau de l'organisation des travaux : une canalisation en béton non visitable peut parfaitement être remplacée aujourd'hui par une canalisation en plastique de même caractéristiques.

Cette implémentation, ajoutée à toutes les précédentes n'as pas permis d'améliorer les estimations des matériaux, car trop d'éléments étaient jugés comme mal labellisés. Cependant, cet outil a permis de justifier le regroupement des classes qui c'était alors fait jusque-là uniquement sur des règles métiers.

#### 2.4.3.2.4. Soft-voting

Pour l'estimation des matériaux, malgré la pondération et la synthétisation de données, le déséquilibre des classes reste un problème. Une autre solution mise en place pour le limiter est le soft-voting. Plutôt que d'avoir un modèle aléatoire issu de la validation croisée ou de ré-entrainer un modèle pour estimer les paramètres manquants, tous les estimateurs de la validation croisée sont utilisés. Les résultats sont des vecteurs de probabilité d'appartenance à chaque classe et le maximum de ces vecteurs est utilisé pour prédire la classe. À la manière du RandomForest, les apprenants « votent » pour élire la classe la plus probable. Cette étape a permis de limiter les temps de calculs, car les modèles étaient déjà calculés. Il faut cependant faire attention avec cette méthode. Elle est valable si le jeu de donnée ne possède pas de biais. En effet un classifieur avec un jeu d'entraînement ultra-spécifique pourra donner des résultats biaisés (ne prédire que du Béton armé à 100 % par exemple). Il a été vérifié que cela n'était pas le cas lors de la phase de développement et de test et la moyenne des écarts-types des précisions des classifieurs est de 0.003, nombre nous confortant dans l'utilisation du soft-voting.

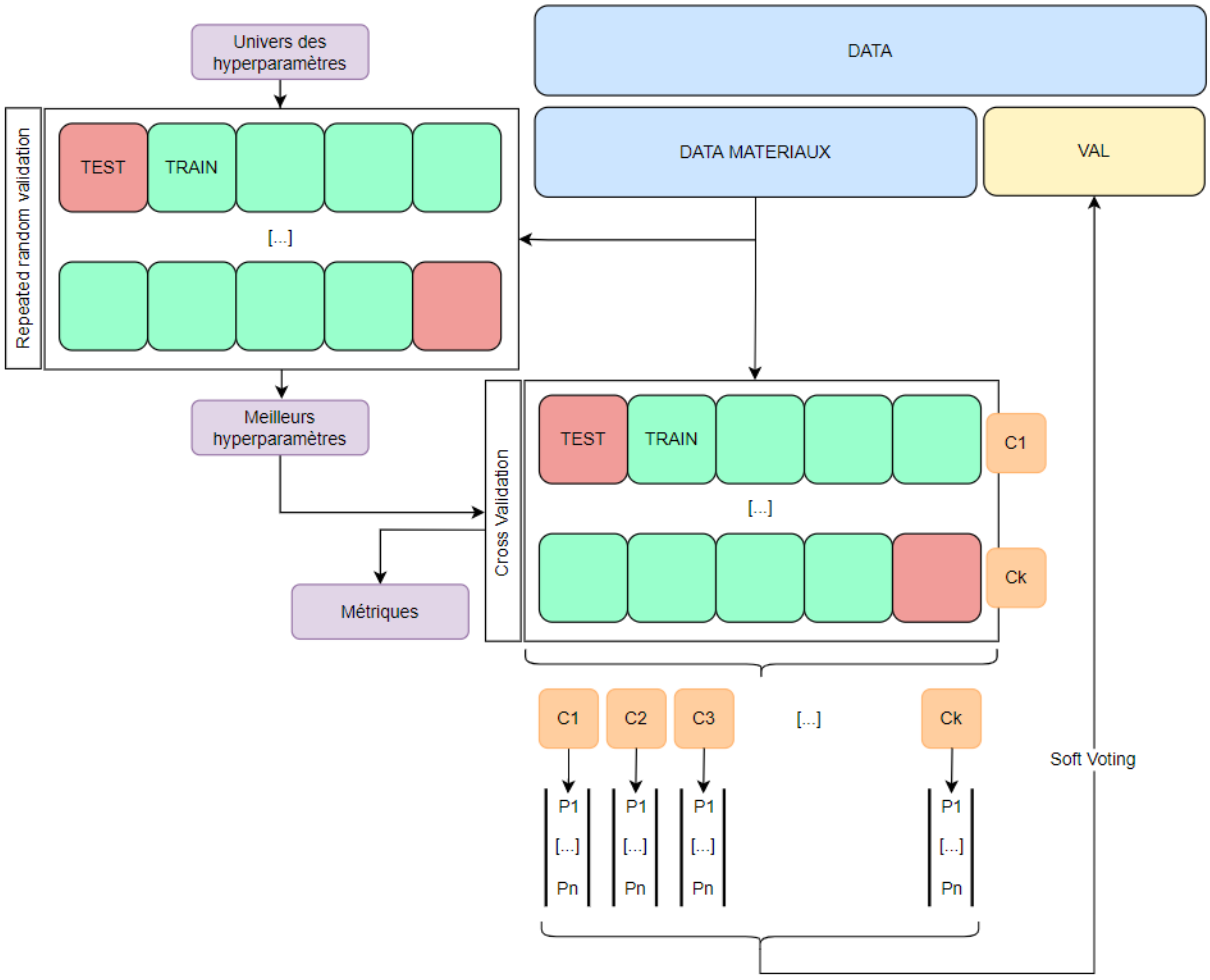



*Figure 17 Processus final d'estimation des matériaux avec soft-voting*

Cette amélioration couplée à celles commune aux années de pose, à la pondération et à la synthétisation de données, a permis d'augmenter le F-score de 58.5 % à 84.7 % par rapport aux travaux de 2020-2021 et d'apporter des résultats de facto beaucoup plus stables car soumis à soft-voting.

#### 2.4.3.3. Résultats et discussions

L'estimation des matériaux par apprentissage supervisé permet de passer d'une connaissance de 58.8 % à 100 %. Les matériaux estimés ont pour qualification une précision, un rappel et un f-score – moyenne harmonique de la précision et du rappel.

| Xgboost + CV + Pondération + Oversampling post CV (SMOTE) + Soft-vote 7 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Classe | Précision | Rappel | F-Score | Moyenne | Précision | Rappel | F-Score |
| AMCI | 0,859 | 0,897 | 0,877 | Pondéré | 0,893 | 0,884 | 0,886 |
| AUTR | 0,707 | 0,847 | 0,771 | Macro | 0.826 | 0.872 | 0.847 |
| BTAM | 0,955 | 0,879 | 0,915 | | | | |
| BTAU | 0,875 | 0,909 | 0,892 | | | | |
| FON | 0,820 | 0,794 | 0,807 | | | | |
| PLAS | 0,748 | 0,892 | 0,813 | | | | |
| ROCH | 0,820 | 0,890 | 0,853 | | | | |

*Tableau 5 Résultats d'estimation des matériaux*

D'un point de vue général (Tableau 5), le modèle actuel a une précision de 0.826 et un rappel de 0.872. C'est-à-dire que la proportion d'éléments corrects parmi ceux retournés pour une classe est de 82.6 % et que la proportion d'éléments corrects retournés parmi ceux qui existent est de 87.2 %. Toutes les métriques propres aux matériaux sont au-dessus des 70 % et sont supérieures à leurs homologues des travaux précédents.

Si l'on s'intéresse à la distribution des prédictions dans les différentes classes de matériaux, plusieurs résultats sont intéressants. Premièrement, il est observable sur le graphique ci-dessous (Figure 18) que les estimations précédentes de 2020-2021 subissaient très fortement l'équilibrage des classes avec en grande majorité des prédictions pour les classes BTAM, PLAS et BTAU. Il faut noter que les estimations sorties par le modèle d'apprentissage supervisé ne suivent pas la proportion d'éléments des classes connues dans les estimations. Ce n'est pas parce que 30 % des matériaux connus sont en classe PLAS que 30 % des inconnus le seront. Cet effet se remarque notamment sur les classes AMCI et FON. Ce sont des matériaux spécifiques qui ont été posés dans des périodes ou avec des contraintes particulières. Les collecteurs en fontes qui se situent généralement sous les ponts ont déjà été en grande partie inventoriés et il est donc normal pour l'algorithme de ne plus trouver autant. Il en est de même pour l'amiante, interdite en France depuis 1997.

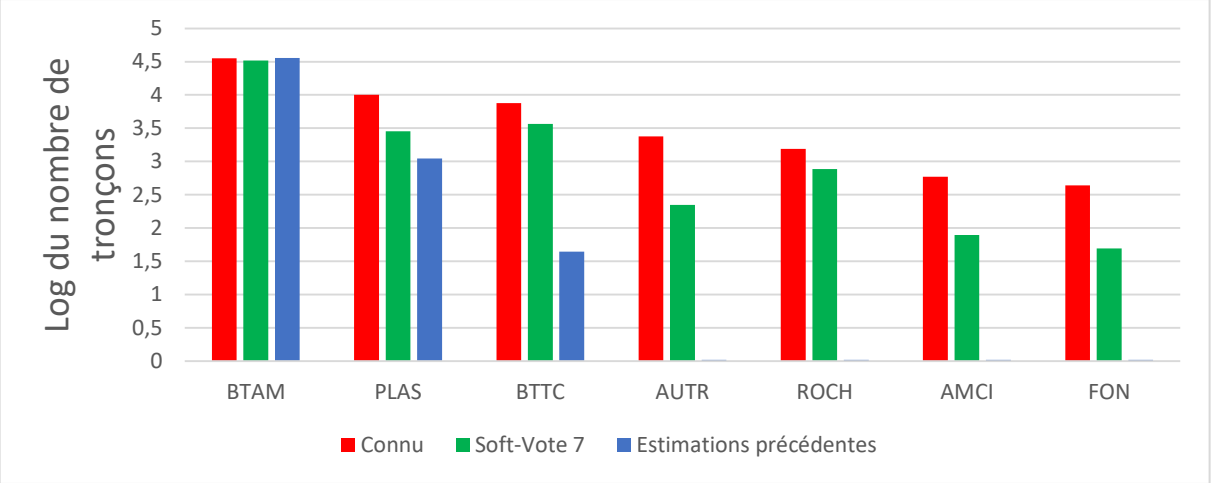



*Figure 18 Comparaison des résultats avec les anciennes estimations*

Le fait que la classe BTAM (majoritaire dans les données d'entraînement) soit la plus représentée dans les données prédites peut signifier deux choses. Soit l'algorithme est encore assujetti à des problèmes de déséquilibre des classes lors de l'apprentissage, l'amenant à surestimer cette classe ; soit la réalité terrain est telle que prédite par l'algorithme. Le béton est un matériau qui représente la majorité des données d'entraînement et à juste titre puisque c'est la réalité du terrain. L'hypothèse la plus probable est l'abondance de la classe BTAM selon les équipes techniques.

Au niveau de l'analyse spatiale, le détail le plus intéressant est le respect des zones géographiques. Dans les images suivantes (Figures 19 et 20) qui sont une extraction des données de la métropole sur la ville de Lyon, on observe le respect des quartiers anciens. Les matériaux inconnus du quartier des pentes de la Croix-Rousse correspondent à la classe ROCH, matériau cohérent avec l'âge du quartier. Les espaces ou le BTAU prédomine est rempli par des matériaux de classe BTAU. Il semble se dégager une règle métier : les éléments les moins linéaires sont ceux qui vont se voir attribuer la plupart du temps la classe plastique ou autre. Les éléments de fontes se retrouvent au bord des rivières ou alors proches d'ouvrages contenant déjà des canalisations en fonte. Les collecteurs amiantés eux se retrouvent près de collecteurs déjà connus comme amianté et créent des zones. Cependant, certains sont catégorisés comme amiantés sans que le facteur proximité ait l'air d'avoir eu une influence. Il serait donc intéressant de se rendre sur le terrain sur des collecteurs spécifiques afin d'aller vérifier les estimations de l'algorithme. Cette opération n'a malheureusement pu être organisée durant ce stage.

Ces estimations après avoir été présenté à l'équipe Gestion Du Patrimoine (GDP) ont pu être validées grâce à leurs connaissances du réseau et de sa gestion. Ces données seront donc intégrées dans le SIG sous la forme d'un attribut matériaux estimés.

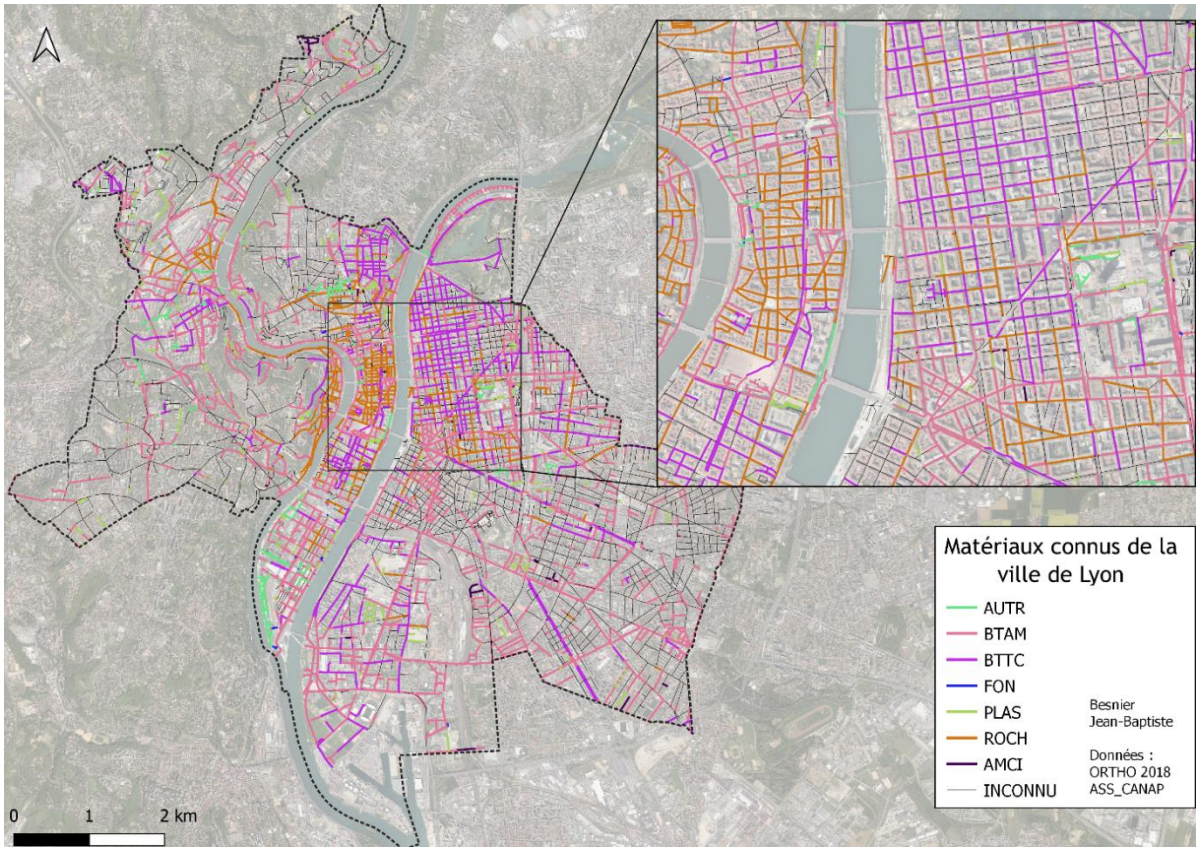



*Figure 19 Export des matériaux connus du 22/04/2022*

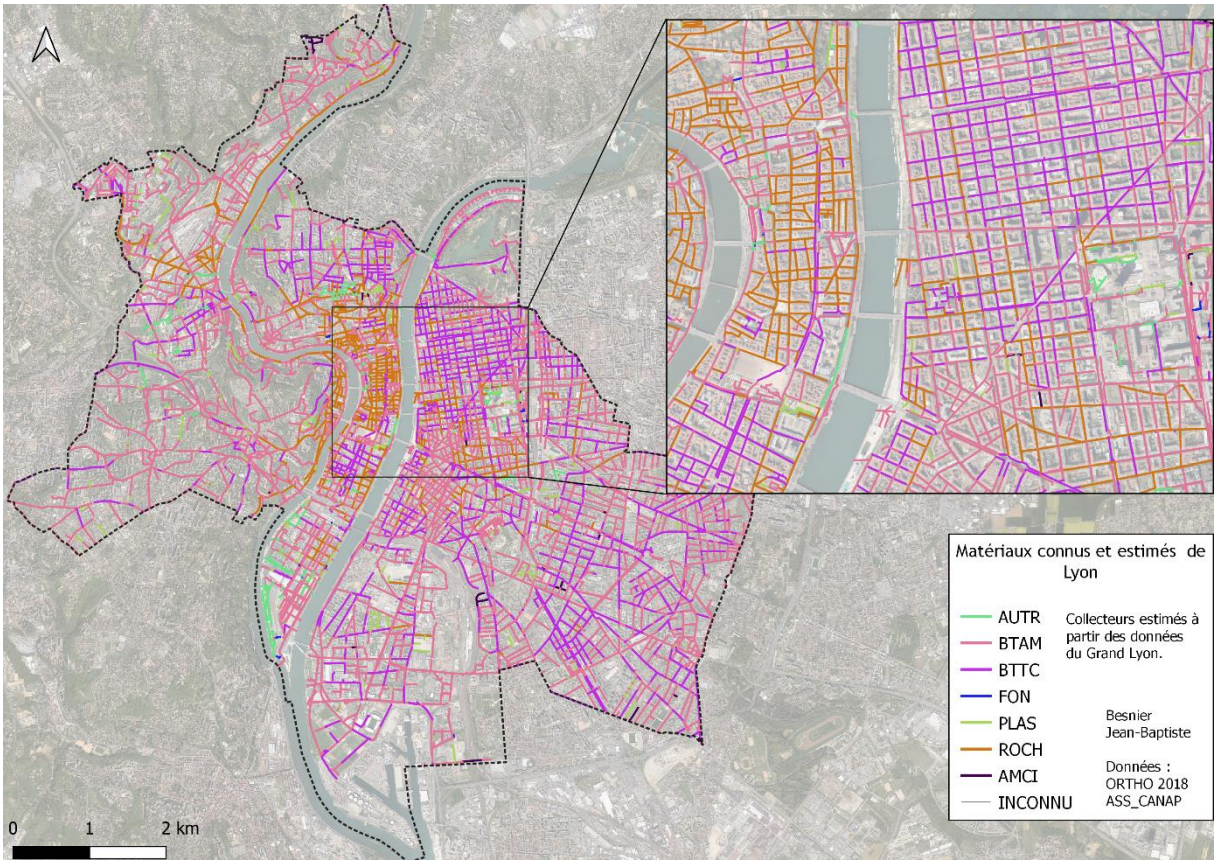



*Figure 20 Estimation des matériaux par apprentissage machine : F-Score : 84.7%*

### 2.4.4. Optimisation de la recherche de données manquantes

Comme vu dans la partie précédente, le rajout des données de localisation des collecteurs a permis à l'algorithme de spatialiser les données. Cependant, si les résultats doivent encore être améliorés, il faut continuer à alimenter l'algorithme en données. Que l'algorithme fonctionne ou pas, les équipes terrains notamment le service d'exploitation ESX, continueront à faire des interventions. Il est donc intéressant d'un point de vue de recherche de la donnée, de diriger les inspections dans des zones où les matériaux ne sont pas connus. Cela permettra en un minimum d'inspection, de rajouter de l'information sur les matériaux qui est dépendante de la localisation. Si en plein milieu d'un quartier ou d'une rue où tous les matériaux sont inconnus, donner le matériau du collecteur situé au milieu permettrait à l'algorithme d'estimer ces voisins. Cependant demander à l'équipe ESX de visiter les 1517 km de collecteurs aux matériaux inconnus n'est pas chose envisageable.

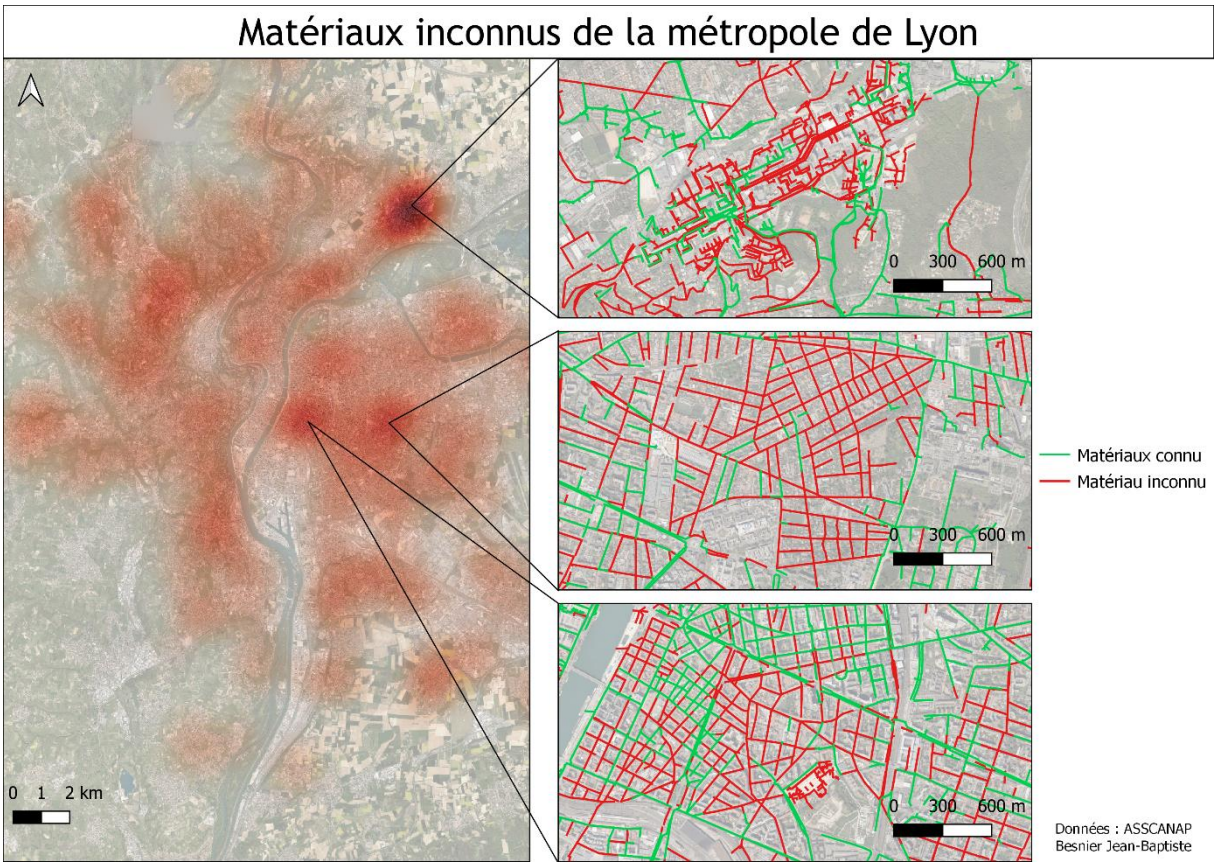



*Figure 21 Repérage des zones à enjeux de la métropole*

Dans un premier temps, une carte de chaleur des collecteurs aux matériaux manquants a été produite (Figure 21). Les îlots de chaleurs sont pondérés par la longueur des tronçons et par leur densité. Cependant, bien que trois quartiers aient pu être ciblés, cette méthode ne permettait pas d'avoir un plan d'action précis des zones à étudier.

Il fallait un moyen pour les équipes terrains d'avoir la visualisation de la priorité des collecteurs à inspecter de manière plus fine. De plus, il n'était pas envisageable de passer tous les collecteurs aux matériaux inconnus en priorité haute d'inspection. Il faut donc sélectionner des collecteurs qui ont potentiellement plus d'informations à ramener pour l'algorithme que les autres.

Pour cela, la couche des collecteurs reliés topologiquement a été utilisée. Cette couche SIG sert lors de modélisation hydrographique et des tronçons qui ne sont pas connectés, mais qui sont dans la continuité du réseau, sont connectés par des collecteurs factices. Un exemple est un regard de visite entre deux tronçons qui laisse un espace entre les deux dans le SIG bien que dans la réalité, l'eau s'écoule de l'un vers l'autre. Le réseau peut être alors vu comme un graphe, les arrêtes sont les tronçons des collecteurs et les sommets sont les intersections et les extrémités de tronçons. Il est alors possible d'utiliser des algorithmes de parcours de graphes, notamment ceux du plugin QGIS Network Analysis Toolbox. En sélectionnant les sommets des tronçons ayant une information de matériaux et en réalisant un parcours de graphe selon l'algorithme de Dijkstra. Il est possible de récupérer la distance la plus courte séparant un collecteur sans matériaux à un collecteur avec matériaux. Ainsi, les tronçons prioritaires sont ceux qui sont situés par parcours de graphe à plus de 200 m d'un tronçon aux matériaux connus.

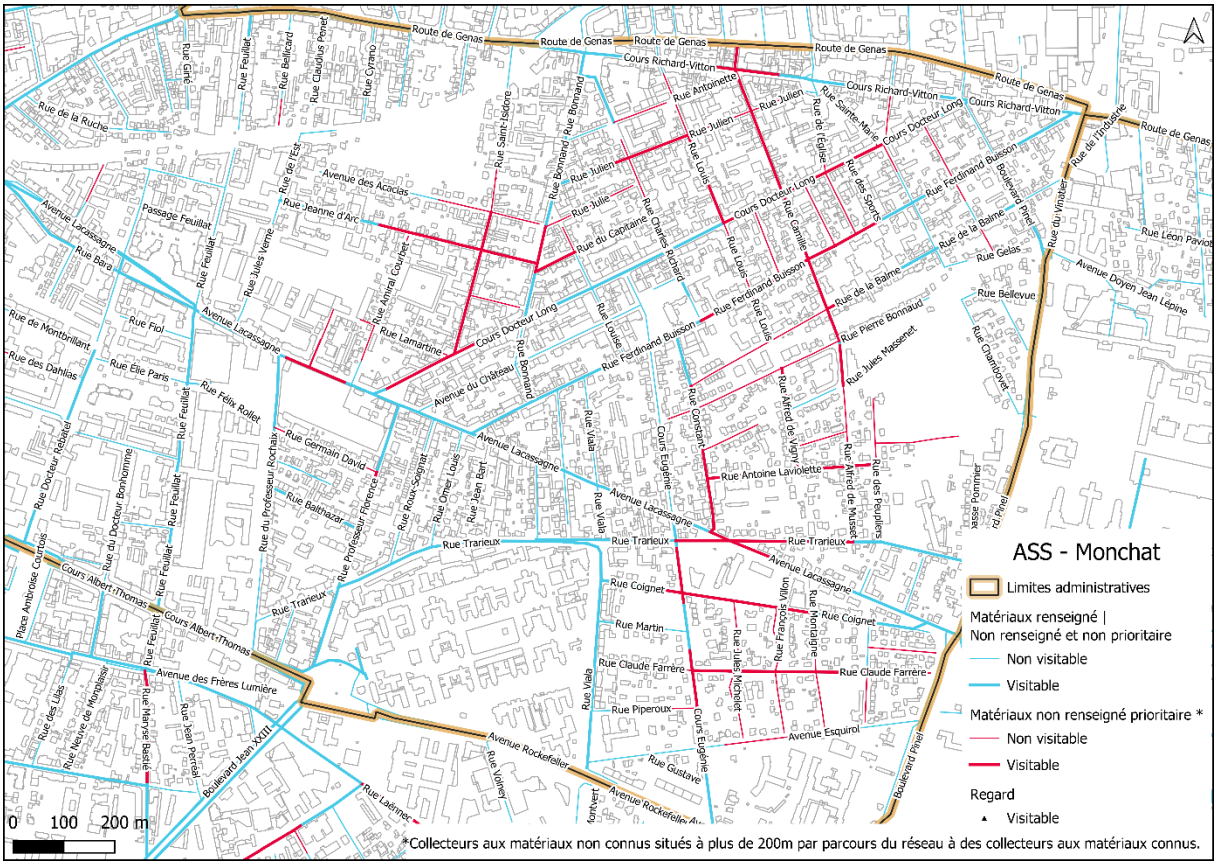



*Figure 22 Carte des collecteurs prioritaire pour la récupération des matériaux - Quartier Monchat*

Cette recherche des collecteurs permet de passer de 1 517 km aux matériaux manquants à 278 km de tronçons prioritaires. Des cartes (Figure 22) et un projet cartographique ont été laissés pour que les équipes d'exploitations puissent les intégrer dans leurs interventions.

278 km de réseau à inspecter reste une quantité de linéaire énorme et représente 6 562 collecteurs soit 6,6 % du total du nombre de collecteurs. Pour diminuer encore le rayon d'action, il faut se focaliser sur les trois quartiers de la métropole présentant la plus forte densité de tronçons manquants. Ainsi en allant analyser les 31 km (0,8 % des collecteurs) des quartiers de Montchat, Guillotière et Ville nouvelle Est, il sera possible de remonter des tronçons ayant un fort potentiel dans les algorithmes d'estimations.

## 2.5. Estimation des notes d'état de santé Indigau

### 2.5.1. Présentation de l'obtention des notes

Les inspections de collecteurs non-visitables (diamètre < 1 m) se font grâce à des caméras robotisées sur roues qui parcourent les canalisations. Le résultat de cette inspection télévisée est un rapport PDF présentant les aléas rencontrés lors de l'inspection. Un dossier avec les vidéos et les captures d'écran est également disponible.

L'entreprise Altereo, prestataire de la métropole de Lyon, a fourni un outil pour estimer l'état d'un collecteur. À partir de ces informations d'inspection, un code d'état de santé est donné à chaque canalisation inspectée. Cette note appelée score Indigau va de G1 à G4 en fonction de la gravité de dégradation du collecteur. Cette note est calculée à partir d'informations intrinsèques au collecteur comme des fissures, des dépôts, des déformations ou encore des branchements défectueux. Il est intéressant de voir si ce score peut être estimé à partir de données extrinsèques au collecteur notamment les matériaux et les années de pose.

| Code | G1 | G2 | G3 | G4 |
|---|---|---|---|---|
| Support | 11260 | 2857 | 2038 | 795 |

*Tableau 6 Nombre d'éléments dans le jeu d'entrainement*

Sur les 79 445 collecteurs non-visitables où une ITV peut éventuellement être effectuée, 16 950 collecteurs ont déjà une note Indigau, soit 21,3 % du total. De manière encore plus accentuée que les matériaux, la note G1 (collecteur en bonne santé) est prédominante dans le jeu de données (Tableau 6). Afin d'estimer les notes Indigau, les améliorations faites aux années de pose et aux matériaux ainsi que les données estimées ont été utilisés.

### 2.5.2. Résultats

En plus des améliorations déjà présentées, la profondeur des arbres a été augmentée successivement dans le modèle afin d'essayer de construire des résultats plus complexes, au risque de sur-apprentissage.

Les résultats obtenus ne sont malheureusement pas à la hauteur de données utilisables (Tableau 7). L'algorithme d'estimation fonctionne, mais n'est pas probant. La très grande majorité des collecteurs estimés est classée en G1, classe majoritaire du jeu d'entraînement.

| | Précision | Rappel | F-Score | Moyenne | Précision | Rappel | F-Score |
|---|---|---|---|---|---|---|---|
| G1 | 0,822 | 0,895 | 0,857 | Pondéré | 0,689 | 0,714 | 0,699 |
| G2 | 0,509 | 0,456 | 0,481 | Macro | 0.551 | 0.518 | 0.534 |
| G3 | 0,391 | 0,325 | 0,355 | | | | |
| G4 | 0,484 | 0,395 | 0,435 | | | | |

*Tableau 7 Estimation des notes Indigau en quatre classes*

Avec les données fournies en entrée, il n'est pas possible d'estimer la note Indigau. C'était un résultat qui pouvait être attendu puisque qu'aucune donnée en entrée de l'algorithme d'estimation ne sert à calculer la note Indigau, basé sur les aléas remontés par les ITV. De plus, l'ajout de la localisation des collecteurs jouait un fort rôle dans l'estimation des matériaux et des années de pose.

Ici et comme visible dans l'image ci-dessous (Figure 23), il y a une très forte hétérogénéité entre des collecteurs se suivant. Ils sont quasiment identiques en caractéristiques (matériaux, profondeur, pente, localisation, etc.) et apportent la preuve que les données externes actuelles ne permettent pas d'identifier l'état de santé tel que défini par Indigau.


*Figure 23 Non uniformité des codes Indigau le long d'une canalisation*

Néanmoins, pour essayer d'avoir des résultats exploitables, un regroupement en deux classes a été effectué pour catégoriser les collecteurs en mauvais état de santé. Mais là encore, la classe la plus intéressante G4 possède un F-score de 0.49 qui n'est pas exploitable (Tableau 8).

| | Précision | Rappel | F-Score | Moyenne | Précision | Rappel | F-Score |
|---|---|---|---|---|---|---|---|
| G1 | 0,877 | 0,915 | 0,896 | Pondéré | 0,816 | 0,823 | 0,82 |
| G4 | 0,551 | 0,447 | 0,494 | Macro | 0,715 | 0,682 | 0,695 |

*Tableau 8 Estimation des notes Indigau en deux classes*

Après discussion avec Frédéric Cherqui, la boite noire qu'est la note Indigau donne une note en fonction de plusieurs sous-états de santé. Notamment un état structurel, un état d'infiltration et un état de courbure lié à la pose. L'année de pose estimée est vue comme un bon indicateur de l'état de santé structurel d'une canalisation. Cependant, les autres états de santé pondérant la note finale ne sont pas déductibles des informations du SIG et donc il est très compliqué d'estimer les notes à partir de ces informations.

## Transition

Les objectifs visant à améliorer la connaissance des matériaux et des années de pose des collecteurs de la métropole de Lyon ont été rempli. Grâce à l'apprentissage supervisé, ces informations, qui seront remontées dans la base de données sont complètes à 100 %. Cependant, il est important de rappeler que l'apprentissage permet d'obtenir des estimations à partir de données connues. Pour améliorer les connaissances sur les matériaux, il faut remonter dans la base de données des informations de qualité qui n'ont pas été exploitées. Améliorer la précision sur les classes des matériaux pourrait amener à augmenter la précision sur les années de pose. De même avec seulement 30,4 % de données connues pour les années de pose, réussir à remonter de l'information permettra de réaliser des estimations plus précises. Dans la partie suivante, il sera présenté une méthode pour récupérer des informations sur les années de pose à partir du réseau de collecteurs.

# 3. Estimation d'informations attributaires par propagation

À partir des attributs des collecteurs, de leurs géométries et des connections qu'ils forment au sein du réseau d'assainissement, il s'agit de déduire des années de pose de collecteurs à partir des années de pose connus de leurs voisins. Cette étape, possède deux avantages. Le premier est de remonter de l'information fiable dans le SIG à partir de règles métier. Le deuxième est que cet ajout de données ne peut qu'améliorer les modèles déjà existants sur les estimations des années de pose.

## 3.1. Présentation

### 3.1.1. Travaux précédents

Les travaux de 2019-2020 ont abouti à une méthode de récupération des années de pose. Il s'agit d'analyser les collecteurs voisins de collecteurs disposant d'une année de pose. Si les voisins ne possèdent pas d'année de pose et s'ils partagent les mêmes caractéristiques prédéfinies alors l'année de pose est propagée. Il apparaît que cet algorithme fonctionnerait en récursif : une fois cette nouvelle canalisation datée, elle peut être ajoutée à l'ensemble des canalisations datés et aider à dater ses voisines. Cet algorithme a été développé en deux versions. La première, servant de preuve de fonctionnement, recherche simplement les voisins non datés des canalisations datés et propage l'information d'année de pose. Le deuxième en revanche, teste la cohérence entre plusieurs attributs avant de propager les informations d'années de pose. Il se base sur le type du réseau, le type d'effluent, la forme et le diamètre de la canalisation ainsi que sur le matériau utilisé pour propager ou non l'information. Les résultats obtenus sont les suivant, pour le premier algorithme, celui sans contrainte de propagation, au bout de 6 passages récursifs, s'arrête et a propagé 6 034 années de pose. Pour le second, qui propage l'information avec contrainte, 492 collecteurs se sont vus rajouter une année de pose au bout de seulement 3 passages. L'auteur statue sur ces résultats par : « Comme nous pouvions le redouter, le gain en années est extrêmement faible et ne permet pas d'influer sur la connaissance temporelle ». En effet, seuls les 492 tronçons remontés appartiennent très certainement aux travaux de réalisation des canalisations voisines.

Il apparaît ici des incohérences entre la méthode expliqué et les résultats obtenus. Le premier algorithme a réussi à remonter 6,27 % d'information supplémentaire. Mais au vu de ce qui est expliqué dans les travaux, normalement, toute canalisation connectée sur le même sous-graphe qu'une canalisation possédant une année de pose devrait être peuplé. Une estimation des résultats attendus a alors été faite. Dans la méthode présentée, une recherche de voisinage est faite à 3 m autour des extrémités des collecteurs. Pour modéliser cette recherche, des tampons de 3 m autour des géométries des collecteurs ont été construits pour au final extraire 587 sous-réseaux du réseau d'assainissement (Figure 24).

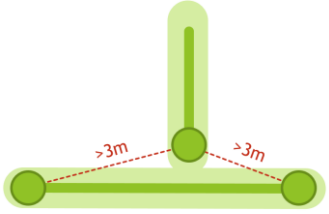



*Figure 24 Approximation du voisinage par tampons*

Parmi ces sous-réseaux, 435 ne possédaient aucun collecteur avec une année de pose. Ils représentent 1 354 collecteurs. Ce qui veut donc dire que les 66 797 collecteurs restants auraient dû se voir attribuer une année de pose. Cet écart entre le nombre de résultats attendus et les résultats publiés s'explique dans l'implémentation de la méthode. Après analyse de celle-ci, il apparaît que la récursivité du programme n'a pas été implémentée. Les boucles ne sont en fait que les voisins successifs d'une canalisation contenant une année de pose.

De plus, la recherche de voisinage ne trouve pour chaque extrémité que le voisin le plus proche ou alors les voisins les plus proches à égale distance. Dans la figure 25 ci-dessous, qui est une représentation des étapes de recherche, bien qu'un collecteur soit dans le rayon de 3 m de recherche, comme il existe des voisins plus proches, il ne sera jamais trouvé. L'algorithme ne fonctionnant pas en récursif, trouvera et modifiera les voisins directs à chaque itération de la même manière.

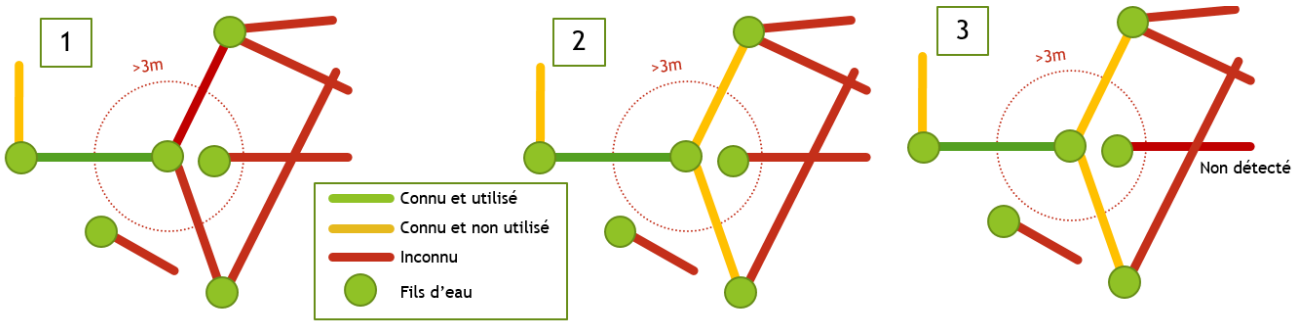



*Figure 25 Visualisation de l'implémentation des travaux 2019-2020*

Cette même implémentation de la recherche de voisinage a été utilisée pour l'algorithme avec contrainte sur la propagation. Cela explique donc le nombre si faible de collecteur retrouvé. De plus, pour le test sur les attributs, trois boucles ont été implémentées pour gérer les conditions de propagation, la première teste la correspondance entre le type de réseau, le type d'effluent, la forme de la canalisation et le type de matériau. La deuxième est comme la première, le test des matériaux en moins. Enfin, la troisième, est identique à la première. Les conditions de la deuxième boucle sont alors incluses dans celle de la troisième. Aucun collecteur ne passera la troisième boucle puisque l'information aura déjà été propagée à la deuxième.

Malgré les problèmes liés à l'implémentation, la base de la méthode théorique est intéressante, car un algorithme de propagation d'information sous contrainte peut permettre de remonter de l'information qui peut être qualifiée de fiable. Si la canalisation en amont a exactement les mêmes caractéristiques que celle en aval alors elles ont dû être posées en même temps.

### 3.1.2. Objectifs et limites

Au vu de la méthode présentée, il y a lieu de ré-implémenter en entier les deux algorithmes, le premier pour valider la méthode de récursivité et le second avec la propagation sous contrainte amélioré. Il est possible de tester simplement la correspondance avec les attributs utiles. Un autre élément doit être ajouté, l'angle entre la canalisation candidate à l'ajout de pause et la canalisation de base possédant l'information. En effet, dans les informations remontées par l'implémentation précédente, des canalisations transversales ont vu une information ajoutée. Cependant, il est d'usage de ne modifier lors de travaux que les tronçons d'une même rue, afin de rajouter de l'information fiable dans la base de données, il faudra filtrer l'ajout des canalisations selon les angles formés.

L'objectif est donc de récupérer des informations fiables afin de venir nourrir l'algorithme de machine learning. On peut alors qualifier de la donnée et augmenter la précision des estimations en même temps. La limite étant qu'avec ce type d'algorithme, il sera impossible d'estimer l'entièreté du jeu de données.

Cette méthode sera implémentée dans le logiciel FME qui est un outil de traitement de données puissant. Je me suis formé tout au long de ce stage sur cet outil en suivant deux formations en autodidacte.

## 3.2. Mise en œuvre

### 3.2.1. Fonctionnement

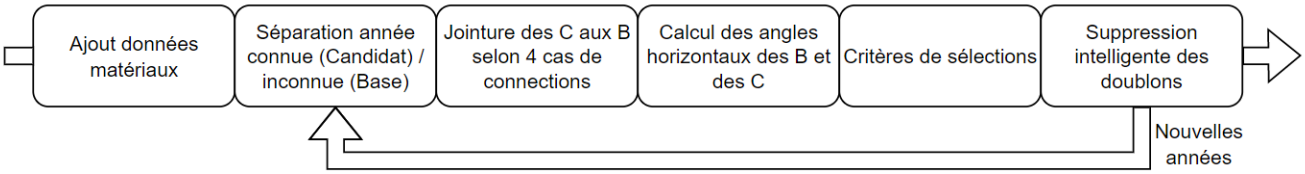



*Figure 26 Synthèse de fonctionnement de l'algorithme de propagation*

La méthode (Figure 26) peut être divisée en plusieurs blocs. Le premier, va ajouter aux attributs des collecteurs les matériaux estimés avec les algorithmes d'estimations. Cela permet de rajouter un critère supplémentaire à la conformité des canalisations pour propager de l'information. Le second bloc va séparer le jeu de données en deux sous-jeux. Les canalisations datées sont les candidats (à utiliser ou non) auxquels seront rattachées les canalisations (Bases) voisines qui ne possèdent pas d'années de pose. Une fois que de nouvelles canalisations auront été trouvées, cette partie et toutes les suivantes tourneront de manière récursive jusqu'à ce qu'il n'y ait plus de canalisation correspondant aux critères de propagation.

Le troisième bloc est une approche différente des travaux précédents : plutôt que de chercher les canalisations dans un voisinage d'un candidat, les bases seront jointes aux candidats par leurs identifiants *fils d'eaux*. Ce sont des identifiants statuant sur l'amont et l'aval de chaque collecteur. Chaque collecteur en possède donc deux, ce qui laisse quatre cas de figure différents.

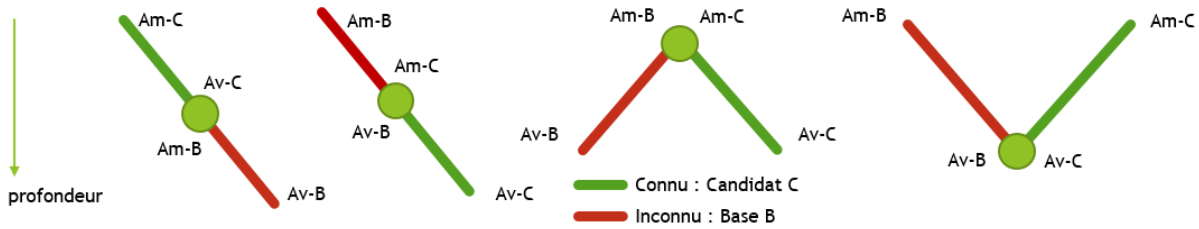



*Figure 27 Cas de figure des connections entre Base et Candidat*

Il est important de représenter et de caractériser chaque type de branchement (Figure 27). Cela aidera notamment au calcul des angles pour le bloc suivant. Le fait de connaître la relation entre deux collecteurs permettra, de proposer des méthodes suivant les différents cas de figure. Un exemple serait de respecter une règle métier qui veut que l'on commence par l'amont d'une rue pour commencer des travaux. Les tronçons en aval ayant donc une année de pose supérieure ou égale à celles connues en amont. Cette règle n'a pas été mise en place dans cette méthode.

Pour le quatrième bloc, l'angle entre la base et les différents candidats doit être calculé. Cependant, cette partie a été très complexe à implémenter (Figure 28). En effet, l'algorithme calculant les angles horizontaux implémenté ne peut se servir que des géométries des collecteurs. Il a fallu trouver un moyen de savoir si le premier nœud d'une ligne était l'amont ou l'aval, selon le cas de figure, inverser les géométries afin de calculer l'angle au bon endroit. Pour cela, les coordonnées des nœuds des géométries des collecteurs et des *points fils d'eau* ont été extraites et la correspondance des nœuds à l'amont ou à l'aval a été faite en étudiant la distance aux points fils d'eau correspondants.

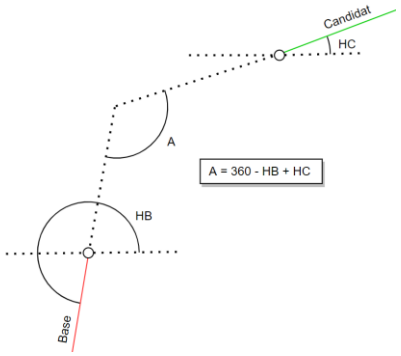



*Figure 28 Angle entre deux canalisations non connectées géométriquement*

Ensuite, le cinquième bloc est le bloc de test qui va s'assurer que la base et les candidats ont des caractéristiques similaires. Le test s'assure également que la canalisation de base est située dans un cône de ± 30° des candidats. Si la base est retenue elle passe au bloc suivant sinon, elle est rejetée. À ce moment du processus, il peut toujours y avoir des canalisations base en plusieurs exemplaires. En effet, si elle possède un candidat en amont, et deux candidats en aval qui respectent tous les critères, il y aura trois entrées pour cette base avec potentiellement trois années de pose différentes. Le dernier bloc s'occupe lui de supprimer les duplicatas. Il a été décidé que si plusieurs dates sont rencontrés et jugées valide pour un collecteur de garder la plus ancienne. Ceci à juste titre, car des canalisations sont régulièrement remplacées par des tuyaux aux caractéristiques identiques, mais à des dates postérieures. L'objectif de ces travaux étant d'aboutir à une meilleure gestion des risques liés aux états de santé, garder la date la plus récente pourrait minimiser ces risques.

Une fois tous les blocs passés, les nouvelles canalisations repassent dans le deuxième bloc pour continuer la récursivité (Figure 29).

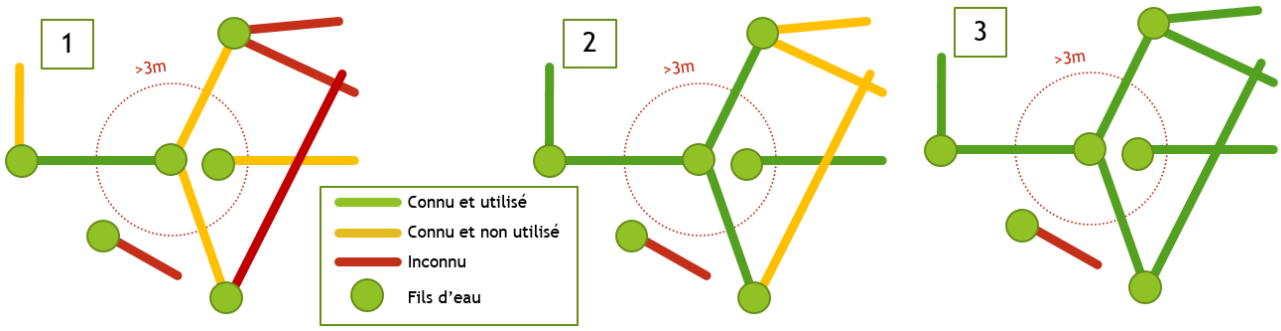



*Figure 29 Fonctionnement de l'algorithme de propagation*

### 3.2.2. Résultats et discussion

Le premier algorithme où la propagation se faisait sans contrainte a permis de valider la méthode (Figure 30). 58 881 collecteurs ont ainsi vu leur année propagée, correspondant à l'ordre de grandeur prévu. Bien que cette méthode n'apporte pas d'information fiable, elle est le squelette de la suivante. Un résultat intéressant cependant est le nombre de collecteurs récupéré à chaque itération de propagation. La courbe (Figure 31) décroît de manière quartique donnant des informations sur la répartition des informations manquantes. Cela montre également que 40 % des collecteurs aux dates inconnues sont à moins de 5 collecteurs d'un collecteur connu et justifie donc l'intérêt d'une méthode par propagation de proches en proches.

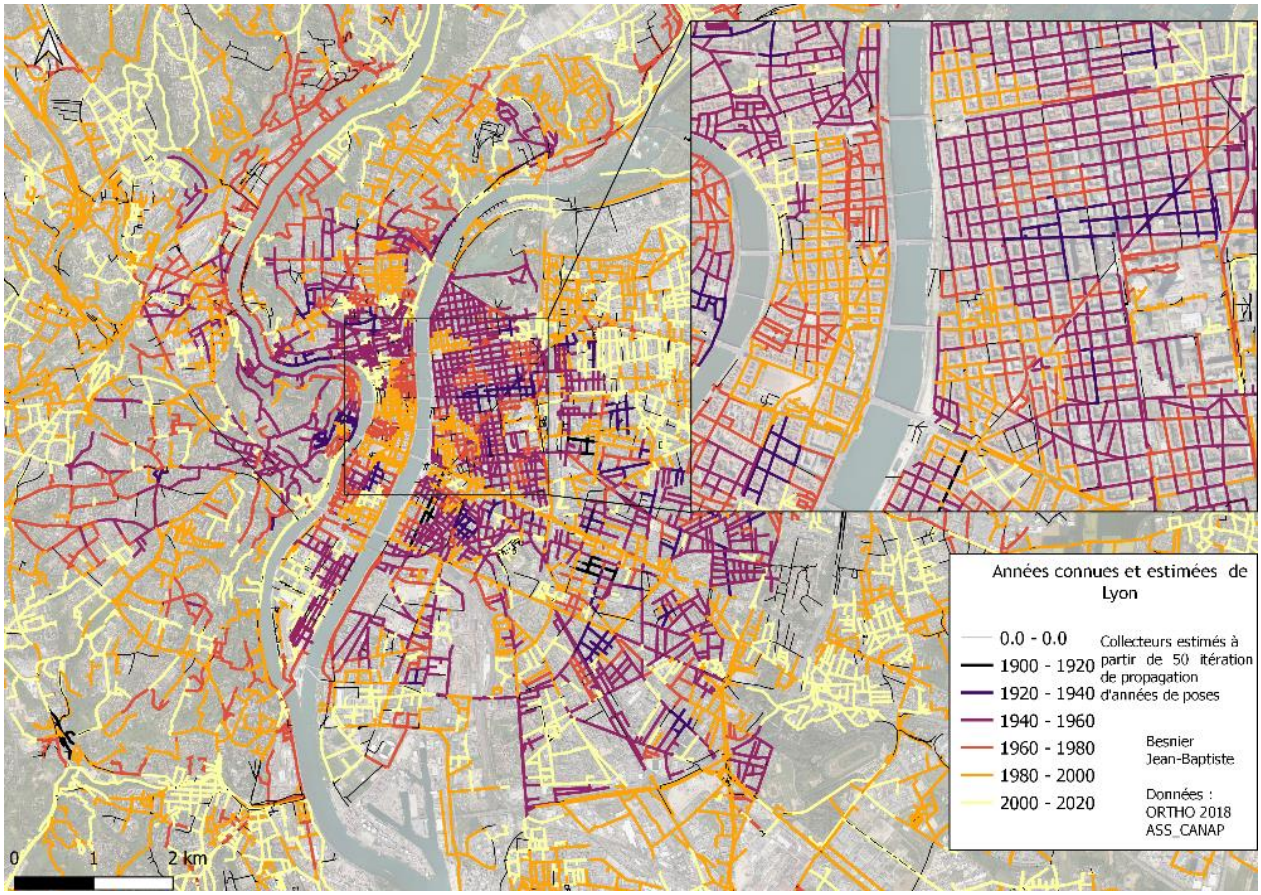



*Figure 30 Années de pose propagées de manière naïve : justification de la méthode*

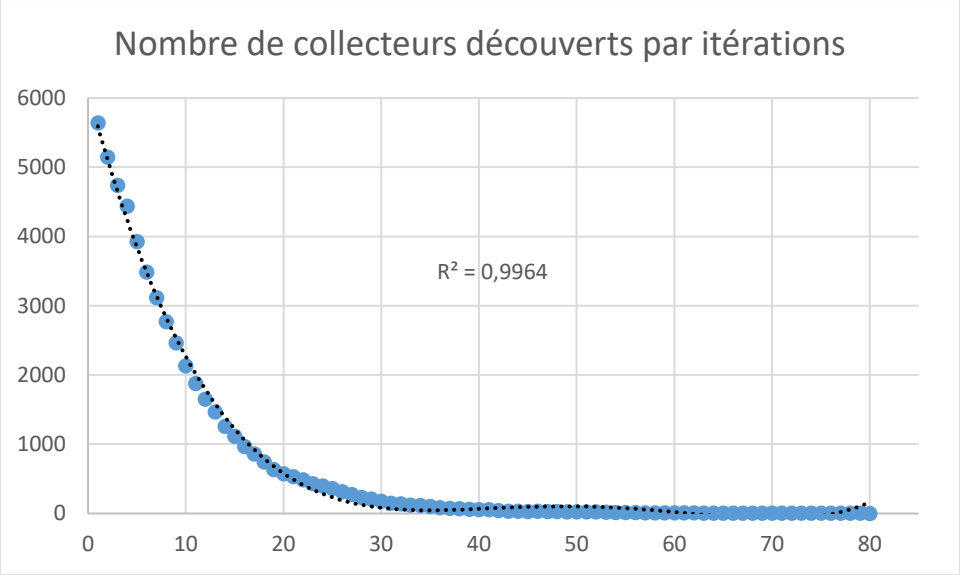



*Figure 31 Étude de la découverte du nombre de collecteurs*

La seconde méthode (Figure 32) renvoie quant à elle 2 648 collecteurs. Ces collecteurs ont répondu à tous les critères de sélection et forment donc un ajout solide et de qualité à la base de données. Cela représente un gain de 2,7 % de connaissance pour les années de pose et une augmentation de 8,5 % pour le jeu d'entraînement pour les algorithmes d'estimations. Ces données supplémentaires ont permis d'améliorer le RMSE sur les années de pose qui passe de 5,8 ans à 5,5 ans.

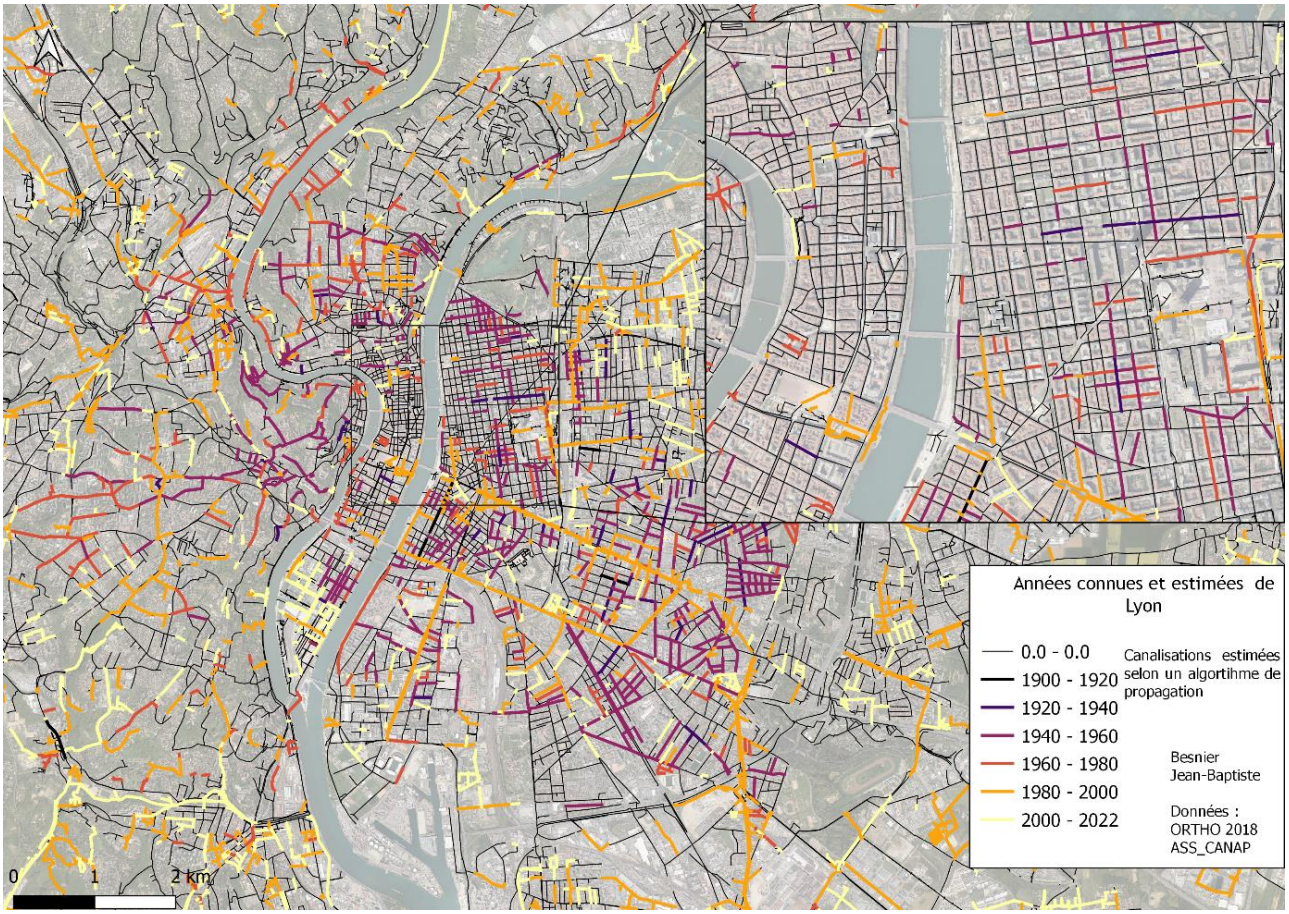



*Figure 32 Ajout de 2 648 collecteurs via propagation sous contrainte*

Cependant, plusieurs limites à ces déductions sont à présenter. Premièrement, si des travaux de réhabilitation sont effectués dans une zone où les dates de pose ne sont pas connues, cette année de pose renseignée, va se propager aux autres canalisations si elles ont les mêmes caractéristiques. Il y a matière à améliorer le modèle pour par exemple ne pas propager les années de pose passé une certaine date. Cependant, les années de pose remontées entre 2010 et 2022 sont marginales et ne vont pas influer sur les résultats d'estimations. Par contre, le fait de remonter les années de pose les plus anciennes peut parfois poser problème (Figure 33). À Villeurbanne, des travaux d'archives ont été réalisés et des années de pose très anciennes ont été remontées dans la base de données (1900-1920), sans savoir si de nouveaux travaux ont été réalisés après. Dans les premières estimations par apprentissage supervisé, les équipes techniques ont statué que les canalisations étaient plus anciennes que les estimations de l'algorithme. La propagation des années de pose les plus anciennes a permis d'augmenter leur nombre et de permettre à l'algorithme de vieillir le réseau Villeurbanne. On observe alors un « forçage » du vieillissement notamment à cause du code INSEE qui est une variable macroscopique du réseau. Pour éviter cet effet de forçage de certains arrondissements, le code INSEE a été enlevé des entrées du modèle. C'est avec ces paramètres que les meilleurs résultats sont obtenus avec un RMSE de 5.4. C'est également cette version qui correspond le mieux à ceux qui ont réalisé des travaux d'archives et qui ont une connaissance accrue du réseau.

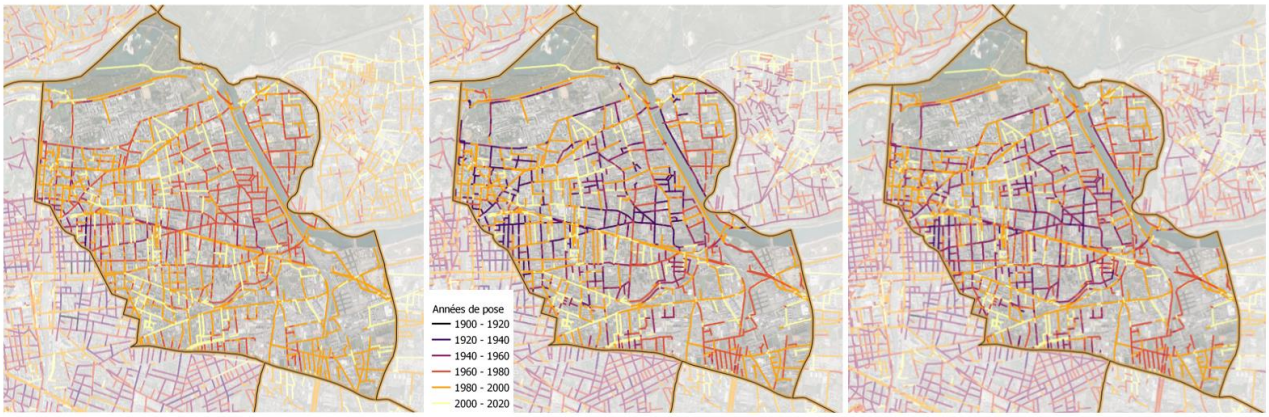



*Figure 33 Importance de choix des données d'entrées : Ancien modèle / Modèle avec propagation/ Suppression code INSEE*

Il faut bien noter que les améliorations sont de plus en plus minimes. L'algorithme possède ses limites et il a fallu choisir les bonnes données en entrées et les bons paramètres pour être fidèle à la connaissance du terrain.

# Transition

Cette partie sur la création d'algorithmes de propagation a été très enrichissante d'un point de vue technique tant sur les logiciels que les méthodes employées. Ces algorithmes n'ont pas encore été implémentés dans un processus d'automatisation par le service géomatique, mais ils pourront être réutilisés chaque année avec les nouvelles informations d'années de pose qui auront été remontées via des travaux d'archives. Ils ont permis d'améliorer la connaissance des années de pose de manière automatique et de participer à l'amélioration des résultats des estimations.

Toujours dans la même démarche de recherche d'information pour nourrir les algorithmes, des discussions ont été mené avec un historien et des techniciens pour remonter de l'information.

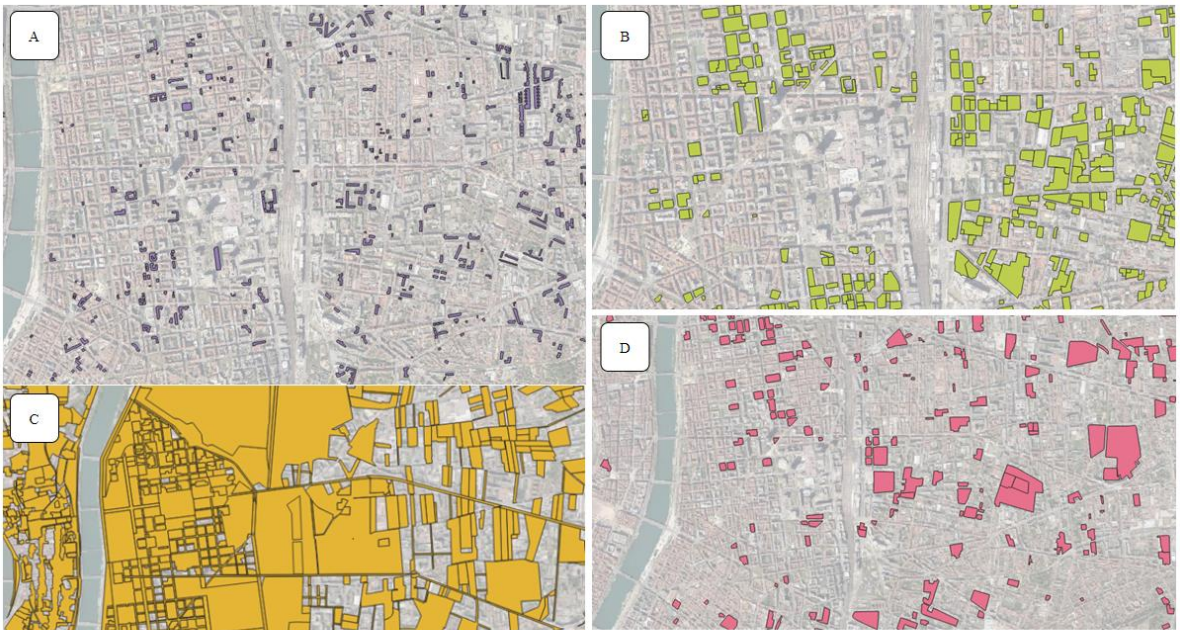
# 4. Informations sur les collecteurs par données d'archives

Après discussion avec Gilles Chuzeville et Aurélie Laplanche responsable équipe GDP réseaux et usines, il est apparu qu'il était intéressant de mutualiser les recherches faites au niveau des archives de la métropole et d'étudier les données non valorisées. Les estimations des matériaux donnent des résultats très satisfaisants et les résultats sur les années de pose ont été améliorés de manière algorithmique. La remontée d'information jusque-là non valorisée dans le SIG présente alors un réel intérêt. Cette partie présente deux manières de remonter de l'information. La première se base sur des travaux d'archives réalisés par d'autres personnes et la deuxième partie porte sur la remontée des informations liées aux inspections télévisées.

## 4.1. Datation par lotissements

### 4.1.1. Présentation des données et méthode

Trois docteurs : Bernard Gauthiez, Nicolas Ferran et Julie Erismann ont travaillé par le passé à la remonté d'informations sur les lotissements de Lyon puis de la métropole. Chapeautée par Bernard Gauthiez, une de leurs missions principales était de réussir à dater la construction des bâtiments ou des zones urbaines afin de retracer l'histoire de la région. De leurs travaux, quatre couches SIG ont été créées contenant des informations différentes (Figure 34).




A. Bâtiment locatif social (construction : 1933-2005)
B. Emprise bâtiment collectif (max : 1950-2003)
C. Lotissements av 1950 (construction : 1416-2016 max 1789-1979 min 1925-2007)
D. Emprise lotissements Erisman J. (construction : 1925-2007 max 1945-2007 min : 1925-2007)

*Figure 34 Données de datations des lotissements*

Pour remonter ces informations, il faut se plonger dans les inventaires de travaux de la métropole de manière manuelle. Une autre méthode qu'ils ont utilisée, est la photo-interprétation. En comparant des images aériennes à des dates différentes, il est possible de donner un bornage des années de construction d'un ouvrage. Les quatre couches contiennent donc soit une année de construction connue, soit une année de construction minimale, soit une année de construction maximale. Les trois docteurs n'étant pas du milieu de la géomatique, les données fournies ont dû être nettoyées.

À partir de ces informations, ils ont essayé de dater le réseau d'eau potable de la métropole. Il ressort qu'à partir des années 1960, les années de constructions peuvent aider à dater le réseau et que la corrélation entre les deux dates est maximale après les années 1990. Il y a donc lieu de se demander si les mêmes conclusions peuvent être faites pour un réseau d'assainissement. Afin d'avoir plusieurs pistes de recherches ou périodes d'études, une étude législative du raccordement des logements à un réseau urbain a été effectuée. La première version du code la santé publique L133 (M) du 26 octobre 1958 statue sur la nécessité d'un raccordement pour les zones urbaines. Il est abrogé le 4 janvier 1992 (L133(Ab)) et le 22 juin 2000 le code de la santé publique L.1331 V1 sort. En 2022, il en est à la version 5. Nous avons donc 4 périodes d'observations, avant 1958 jusqu'à aujourd'hui où aucune loi n'obligeait le raccordement. Après 1958, après 1992 et après 2000.

Pour associer les polygones des lotissements, des bâtiments ou des zones urbaines aux collecteurs, il faut d'abord préparer les données. Par successions de traitements géomatiques, les quatre couches ont été regroupées en une seule. Peu de polygones s'intersectaient, mais quand c'était le cas, la couche avec la plus grande précision était celle utilisée (A et C). La couche D contenant énormément d'erreurs de saisies et de géométries. Une fois ces données réunies, une jointure spatiale par intersection avec les collecteurs a été effectuée. Bien que cette méthode soit simple, elle permet de s'assurer de ne pas joindre de l'information par erreur. Beaucoup de collecteurs sont au milieu des voies et les bâtiments le long de cette voie ne doivent pas avoir d'influence. C'est également la même méthode qui a été mise en place pour estimer les dates du réseau d'eau potable.

### 4.1.2. Résultats

L'intersection entre les bâtiments ayant une année de construction connue et les collecteurs possédant déjà une année de pose a renvoyé 11 508 collecteurs. En séparant les collecteurs selon les périodes définies précédemment, il est possible de montrer la corrélation entre l'année des lotissements et des années des collecteurs. Il est également possible d'estimer une courbe de tendance entre les deux variables.

| Période | Nb collecteurs | Écart moyen | Écart-type | Corrélation | $R^2$ |
|---|---|---|---|---|---|
| 1925-2022 | 11508 | 18 | 20 | 0.090 | 0.009 |
| 1958-2022 | 10292 | 16 | 19 | 0.120 | 0.014 |
| 1992-2022 | 1635 | 3 | 9 | 0.410 | 0.163 |
| 2000-2022 | 949 | 2 | 8 | 0.465 | 0.216 |

*Tableau 9 Synthétisation des résultats de croisement des années*

Le tableau 9 montre ici une amélioration de toutes les valeurs au fil des périodes avec une très nette amélioration à partir de la période post 1992. Cependant, avec une corrélation maximale entre les deux variables de 0,465 et un $R^2$ pour une modélisation linéaire de 0,216, il n'est pas envisageable d'utiliser les informations des lotissements pour dater les canalisations en dessous.

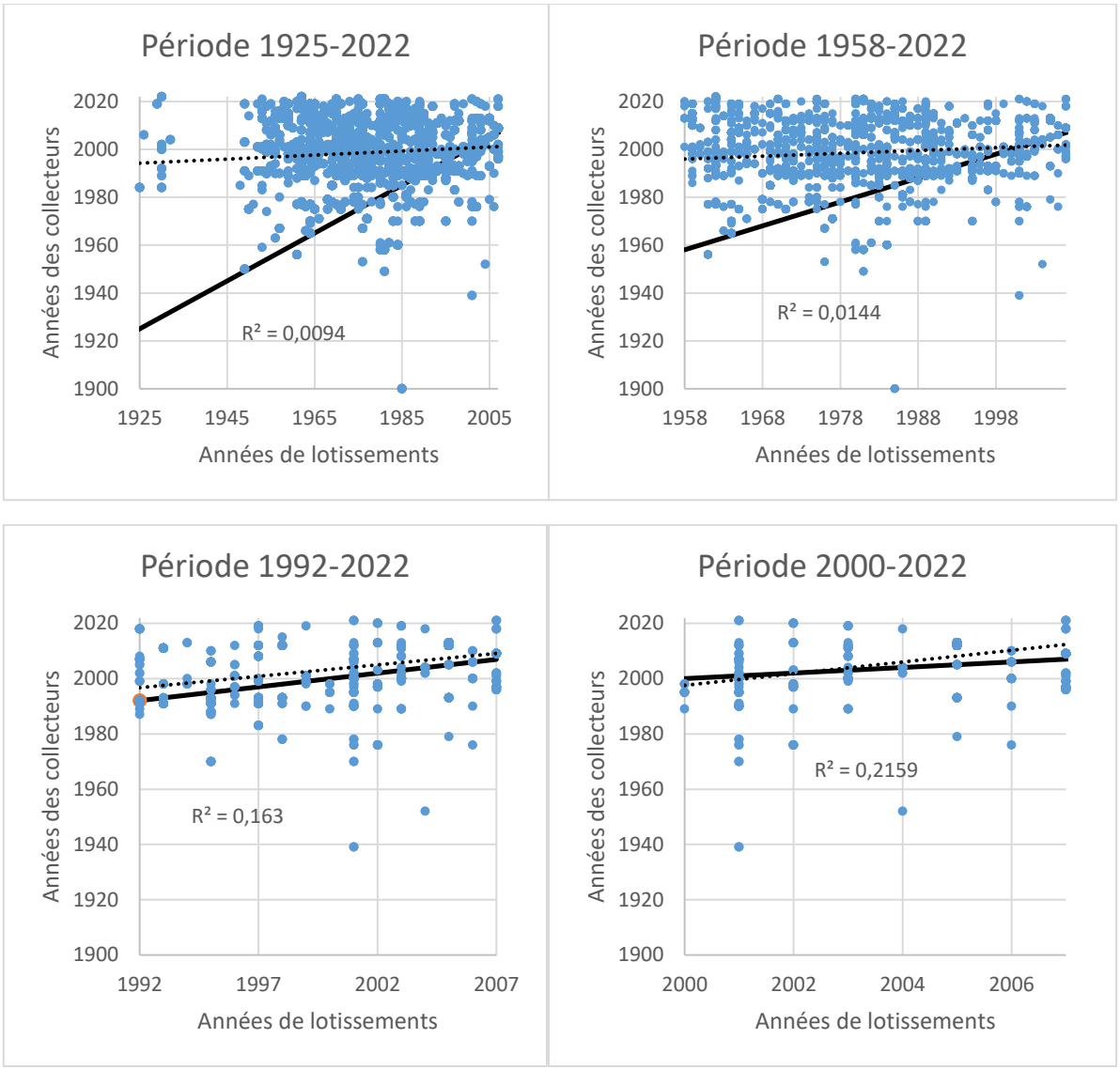



*Figure 35 Résultats et modélisations du croisement années de pose / construction*

Plusieurs explications peuvent justifier la non-concordance des résultats observés (Figure 35). Le fait que tous les résultats s'améliorent peuvent expliquer deux choses. Premièrement, que les collecteurs ont été renouvelés depuis la construction des bâtiments. C'est pour cela que l'écart moyen reste toujours positif. Une autre raison peut-être que la législation n'a pas été mise en place dès sa date de mise en application. Après 1992, il est observé un bon de performance dans les mesures réalisées. Avant 1992, dans la métropole de Lyon, l'article L133(M) n'était pas respecté au niveau des délais de raccordement à l'assainissement. Pour expliquer les collecteurs qui sont plus anciens que les bâtiments en surface, c'est que les collecteurs étaient déjà présents lors de la construction du lotissement qui n'était donc alors qu'une modification d'une zone déjà urbanisée.

## 4.2. Récupération d'informations par inspections télévisées

### 4.2.1. Principe de l'inspection télévisée et données potentiellement récupérables

Des robots téléguidés passent dans les canalisations (D < 1 m) avec une caméra pour faire une inspection d'un tronçon de collecteur. Une inspection se fait soit dans le sens d'écoulement, soit dans le sens contraire à l'écoulement. Le technicien qui contrôle la caméra note un code d'erreur de chaque aléa ou anomalie présent dans la canalisation. À chaque observation, la distance par rapport à l'entrée du collecteur et un angle azimutal positionnant l'observation sur la canalisation est donnée. Toutes ces inspections sont normées et respectent la norme européenne NF EN 13508-2+A1:2011. Elle définit notamment un système de codage pour la description interne des canalisations.

Toutes les ITVs effectuées à la métropole de Lyon depuis 2009 sont stockées sur le serveur (3.5 To) mais n'ont jamais été utilisées comme donnée SIG. Les ITVs sont des outils dont l'utilisation est ponctuelle permettant d'aider à la prise de décision sur un projet en particulier. Pour un projet futur par exemple, impossible de savoir si une ITV a été effectuée en amont ou en aval du projet les cinq dernières années à moins de fouiller tous les fichiers du serveur.

Cette donnée, structurée dans des dossiers par arrondissement/communes, villes et rues n'est aujourd'hui pas remontée dans le SIG. Une information contenue dans ces ITVs est pourtant très intéressante : les matériaux. Dans un premier temps, une méthode a été implémentée pour remonter les matériaux dans le SIG. Cependant, à la fin de ce stage, tous les objectifs ont été réalisés et une nouvelle mission a été engagée. Remonter toutes les informations des ITVs sous forme de couche SIG. Ce projet contenant l'information des matériaux dans un but de synthétisation, seule cette méthode globale sera présentée.

L'objectif est donc le suivant : lire les données des ITVs contenues sur le serveur. Selon le sens d'inspection, remonter le long du ou des collecteurs inspectés sous forme de points, les codes normés des inspections. Si le code d'observation est un branchement, créer une amorce de branchement perpendiculaire à la canalisation selon l'angle azimutal de l'observation. Sur ces branchements, calculer si possible leur hauteur par rapport au fond de la canalisation mère.

Visite d'une réalisation d'une inspection télévisée. Deux opérateurs font descendre une caméra différente en fonction du type (taille, dépôt) de canalisation et un opérateur contrôle la caméra dans le fourgon et note toutes les observations.

### 4.2.2. Exploitations des rapports d'inspections télévisées

Les comptes-rendus d'ITVs sont constitués de trois éléments, un dossier photos et vidéos, un rapport PDF de l'inspection ainsi qu'un fichier TXT servant à générer le PDF. Cependant, pour certains comptes-rendus d'ITVs, seuls les rapports PDF ont été conservés. Pour 7 333 ITVs (7 333 PDF), seuls 5 009 fichiers TXT ont été trouvés. Donc 70,8 % des ITVs sont potentiellement utilisables. Les fichiers PDF n'étant pas fait pour être manipulés, il est très dur d'en retirer de l'information.

Les fichiers TXT se présentent comme suit : une première partie représente l'entête du fichier. Il faut la lire avant le reste de toute autre information puisqu'elle détermine comment lire le fichier. Comme plusieurs canalisations peuvent être inspectées lors d'une seule ITV, les parties d'après peuvent se répéter autant de fois qu'il y a de canalisations. Ainsi, dans ce document, des informations concernant le collecteur sont renseignés. Des sections sur le lieu d'inspection, sur le détail de l'inspection et sur le détail de la canalisation viennent donner toutes les informations sur une inspection. Entre les balises #C et #Z se trouvent alors toutes les observations faites sur la canalisation. Avec à chaque fois, la distance à laquelle est faite l'observation. Il faut au final respecter rigoureusement les normes et nomenclatures et faire attention à tous les champs qui n'ont pas été renseignés.

*Figure 36 Étude de la structure des fichiers de synthèse des ITVs*

Dans l'exemple de fichier de la page précédente (Figure 36), par exemple, le matériau (code ACD section détails de la canalisation) du collecteur 8 292 (code AAA section lieu d'inspection) correspond au code AH. En se référant à la norme européenne, le matériau est « Béton Armé ».

Au vu de la structuration des données en entrée et des résultats attendus, un modèle de données a été décidé (Figure 37). Deux couches seront produites, une couche de géométrie identique aux collecteurs de la couche CANAEXP, comportant toutes les informations relatives à une inspection télévisée pour un collecteur (quatre premières sections hors entête). Une deuxième couche possédant la clef identifiant les collecteurs ayant reçu une ITV, qui contiendra elle, les points avec les informations des observations (dernière section). Puis dans un troisième temps, une couche des amorces de branchement devra être créée à partir des observations d'ITV.

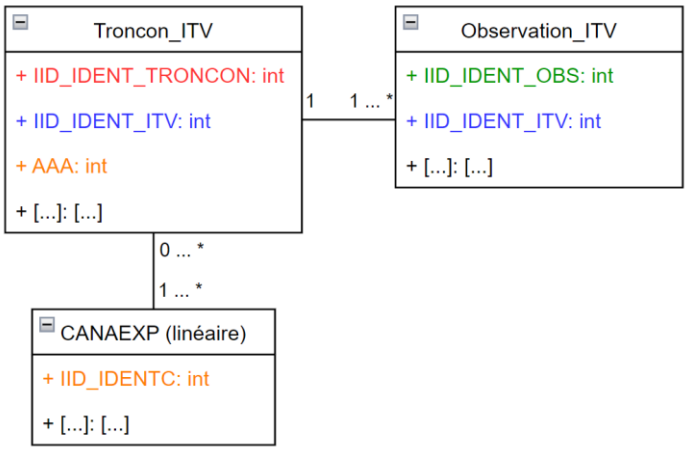



*Figure 37 Schéma de structuration de la donnée des ITVs*

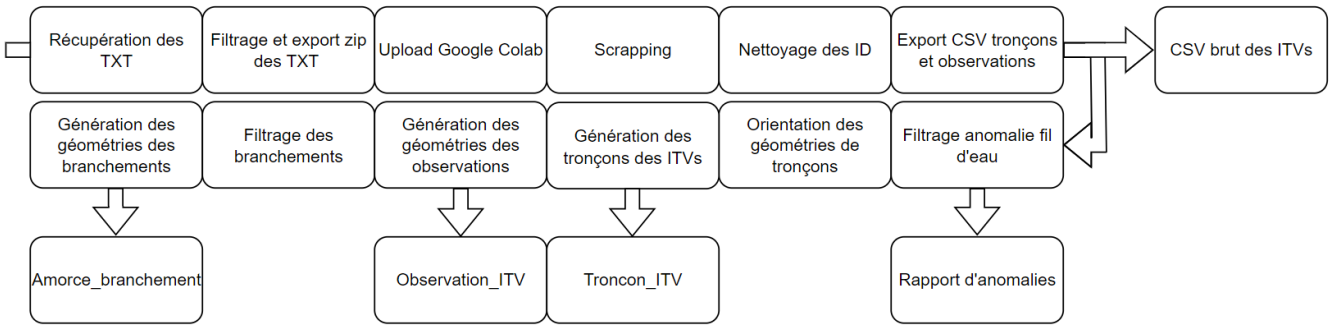



*Figure 38 Synthèse du processus de récupération et valorisation des rapports d'ITVs*

Pour réaliser cela, un premier script FME a été développé pour centraliser l'information (Figure 38). De manière récursive, il va chercher les fichiers TXT des ITVs ayant été validées par le service d'exploitation. En cherchant les fichiers TXT, il faut notamment filtrer des fichiers qui soit ne sont pas des fichiers d'ITV, soit par des fichiers correspondant à la norme NF EN 13508-2+A1:2011 (qui sont marginaux). On passe ainsi de 5 356 fichiers TXT à 5 009 fichiers d'ITVs. Ces fichiers sont ensuite copiés dans un espace de travail pour éviter toute mauvaise manipulation des données, le tout dans une archive compressée.

À partir de cette étape, la lecture des fichiers TXT et l'agrégation de données ont commencés à être implémentés en FME ; seulement plusieurs problèmes sont survenus. FME est un ETL très puissant, cependant, pour lire les fichiers d'ITVs, il faut premièrement lire l'entête du fichier pour connaître tous les paramètres de lecture (délimiteur, marqueur des décimales, etc.) puis relire le fichier avec ces paramètres. FME n'est pas fait pour implémenter simplement des boucles puisque ce sont à chaque fois des appels à d'autres fichiers de projets FME. De plus, le stockage de variable d'un passage de boucle à une autre était compliqué. Il était notamment difficile d'attribuer un ID d'ITV à chaque passage de boucle et de se rappeler quel était le contenu de l'entête pour relire les fichiers. Pour répondre à ces problèmes liés à un manque de compétence dans le logiciel, les traitements de *scraping* des fichiers TXT a été fait en Python. Étant sur un poste de la métropole, Python ne pouvait pas être installé sur le poste de travail. Tout a donc été fait sur Google Colab.

Le travail réalisé sur Python est un travail de structuration de données très classique. Il faut charger le ZIP puis lire les fichiers un par un en construisant les tables des tronçons observés et les observations associées. Les seules subtilités étant l'encodage des fichiers pouvant changer suivant la lecture de l'entête.

Une fois ces deux tables créées, 36 344 tronçons avec potentiellement des doublons, car plusieurs ITV ont pu être réalisées sur ces tronçons ont été trouvés. Ils totalisent 498 655 entrées d'observations. Seulement ces résultats ne sont pas tous exploitables. Les premières ITVs au début des années 2010 ne possédaient pas forcément les identifiants des collecteurs. Ainsi, certains ID renseignés ne sont pas utilisables. Par exemple : « Tronçon n°1 », « Regard de visite 1 > Regard de visite 2 ». Mais certains non-exploitables directement : « RV1>58620 », peuvent être utilisés après traitement. Un traitement a donc été mis en place afin de récupérer le plus d'éléments possibles via des opérations sur les chaînes de caractère des identifiants avec des expressions REGEX et des filtres. Ces opérations ont ensuite permis de joindre 19 331 collecteurs avec 261 423 observations.

À partir de ce moment, il a été possible d'étudier les matériaux remontés par les ITVs et de les comparer aux matériaux connus de la base de données. Cette étude est faite dans la section résultats et discussion. Il reste maintenant à représenter cette donnée. Pour récupérer les géométries des tronçons ayant subi une ITV, la jointure à la base de données permet de récupérer la géométrie. Il faut alors ajouter tous les attributs issus des ITV. Pour pouvoir stocker plus simplement le sens de visite de l'ITV, les géométries ont été mises en accord avec le sens de visite. En effet, pour pouvoir positionner les informations des observations avec uniquement la distance au début de l'inspection, il faut parcourir la géométrie. Si elle est inversée par rapport au sens de visite, les observations le seront aussi.

Pour construire les géométries des observations, une fois la géométrie des tronçons correcte, une ligne est créée le long du tronçon puis est ensuite découpé à la même distance créant ainsi une géométrie ponctuelle. Les informations de l'observation sont ensuite transmises au point (Figure 39).

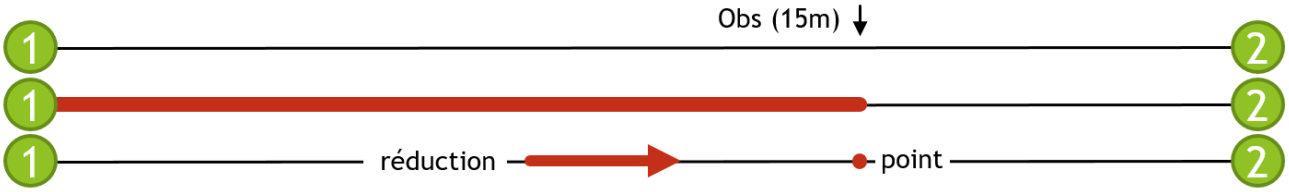



*Figure 39 Étapes de création de la géométrie ponctuelle des observations*

Dans le cadre d'un branchement en revanche, la ligne est créée de la même manière puis laissée à 2 m de longueur pour correspondre à la géométrie d'une amorce de branchement. En fonction de l'information d'angle azimutal lors de l'inspection, une rotation est faite sur l'objet pour qu'il se retrouve perpendiculaire à la canalisation à droite ou à gauche (Figure 40).

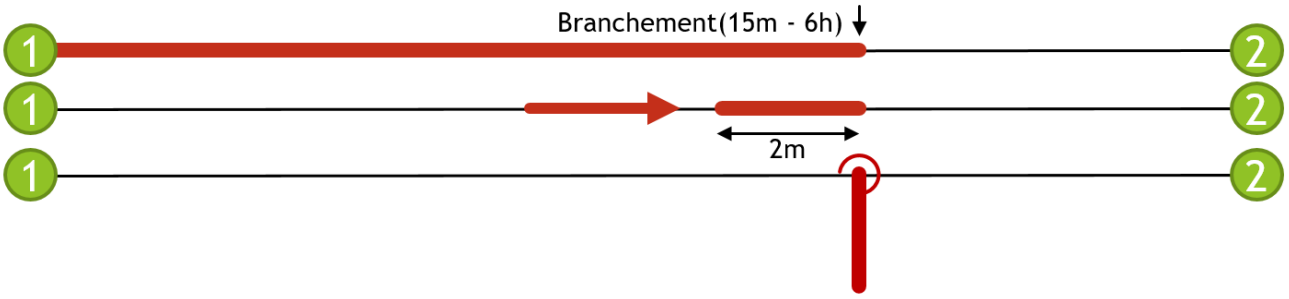



*Figure 40 Étapes de création des branchements*

Pour le calcul de la hauteur (Figure 41), comme cette mission de remontée des ITVs est arrivée à la fin du stage, le temps a commencé à manquer. Récupérer les hauteurs de pose a été fait via des opérations sur les tables sur QGIS, mais n'a pas été intégré au processus complet de récupération des informations des années de pose.

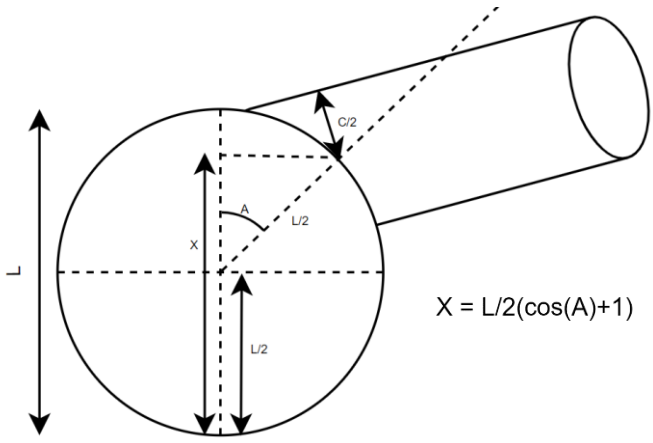



*Figure 41 Estimation de la hauteur de connexion du branchement*

De manière géométrique, il est aisé de rajouter un champ hauteur, connaissant la section de la canalisation inspectée L et l'angle azimutal de l'observation A.

### 4.2.3. Résultats et discussion

La remontée des matériaux était l'objectif principal de l'utilisation des données non valorisées des inspections télévisées. La remontée de toutes les informations et des branchements est arrivé bien après. La première méthode ne concernant que les matériaux n'a pas été présentée car ses résultats sont inclus dans la remontée de toutes les données. Pour utiliser les informations des ITV, trois sources étaient à disposition.

Un fichier Excel qui a été rempli à la main lors du passage des tronçons dans le programme Indigau contient des informations sur les matériaux. Sur 12 176 tronçons joints à la base de données par les mêmes moyens de nettoyage que présentés précédemment, 6 785 ont déjà des matériaux connus dans la base de données. Cela a permis de valider ou non la qualité des données.

Au début des recherches, ne sachant pas quelle donnée qualifiait le mieux la réalité, un indice de Kappa a été calculé (Kassambara A. **[24]**) et vaut 0,35, soit un accord passable (**annexe D** p. 77). Après discussion avec les équipes terrain et l'équipe gestion du patrimoine, il est apparu qu'il était, malgré des formations, très difficile de déduire un matériau. Un béton est indissociable d'un béton armé et un PVC avec du dépôt et de la saleté ressemble à un béton. De plus, la plupart des collecteurs en fonte possèdent un revêtement béton à l'intérieur de la canalisation. Prenant les données de la couche ASSCANAP comme véridiques, il a été possible de calculer les précisions et rappels des classes (Tableau 10). Il a donc été décidé de rajouter les matériaux AMCI (amiante), BTAM (béton armé) et PLAS (plastiques). Le fait que le Béton armé soit remonté, mais pas le béton simple (BTAU) se justifie aussi par le fait que s'il a été précisé qu'il était armé, c'est que le technicien avait des éléments lors de l'inspection pour justifier ce choix. De plus, la classe AUTR dans ASSCANAP peut être vue comme une classe parasite. C'est pour cela qu'il a été décidé, malgré les incertitudes, de remonter les matériaux des ITV quand ils étaient en AUTR dans la base de données. Ces opérations ont permis de remonter 3 865 matériaux soit 3,9 % de données en plus.

| | | ASSCANAP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AMCI | AUTR | BTAM | BTAU | FON | PLAS | ROCH | Somme | Précision |
| ITV | AMCI | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 1,000 |
| | AUTR | 0 | 0 | 5 | 1 | 0 | 19 | 0 | 25 | 0,000 |
| | BTAM | 22 | 89 | 3387 | 217 | 45 | 143 | 2 | 3905 | 0,867 |
| | BTAU | 16 | 67 | 1119 | 126 | 0 | 109 | 3 | 1440 | 0,088 |
| | FON | 0 | 2 | 20 | 0 | 3 | 9 | 0 | 34 | 0,088 |
| | PLAS | 0 | 60 | 334 | 31 | 1 | 861 | 0 | 1287 | 0,669 |
| | ROCH | 0 | 2 | 51 | 4 | 0 | 11 | 21 | 89 | 0,236 |
| | Somme | 43 | 220 | 4916 | 379 | 49 | 1152 | 26 | 6785 | 0,421 |
| | *Rappel* | *0,116* | *0,000* | *0,689* | *0,332* | *0,061* | *0,747* | *0,808* | *0,393* | |

*Tableau 10 Étude de la remontée d'information des ITVs*

Une autre source de donnée était des données historiques de rapports d'inspections de collecteurs visitables. Cette donnée comprenait des codes de matériaux inconnus de la nomenclature et parfois, deux matériaux étaient renseignés. De cette donnée de 3 618 collecteurs n'a été remonté que 26 collecteurs, la grande majorité de ceux connus dans la base de données étant classés en type indifférencié de béton et tombant à 50 % en BTAM et 50 % en BTAU.

La dernière source de donnée était toutes les données contenues sur le serveur. Beaucoup de collecteurs avaient déjà été recensés avec leurs matériaux, car une note Indigau leur avait été attribuée. Cependant, l'analyse de ces données a permis de remonter 1 637 matériaux (+ 1,65 %). Au final 5 528 (+ 5,6 %) informations de matériaux ont pu être ajoutées au SIG. Ces informations ont permis de ré-entraîner les modèle d'estimations et obtenir les résultats du tableau 11

| Xgboost + CV + Pondération + Oversampling post CV (SMOTE) + Data ITV filtrée | | | | | | | |
|---|---|---|---|---|---|---|---|
| index | Précision | Rappel | F-Score | index | Précision | Rappel | F-Score |
| AMCI | 0,943 | 0,890 | 0,915 | Pondéré | 0,924 | 0,924 | 0,924 |
| AUTR | 0,818 | 0,798 | 0,808 | Macro | 0,890 | 0,870 | 0,880 |
| BTAM | 0,952 | 0,947 | 0,949 | | | | |
| BTAU | 0,917 | 0,915 | 0,916 | | | | |
| FON | 0,878 | 0,784 | 0,828 | | | | |
| PLAS | 0,858 | 0,885 | 0,871 | | | | |
| ROCH | 0,865 | 0,875 | 0,869 | | | | |

*Tableau 11 Résultats des estimations des matériaux, améliorée grâce à l'ajout de données filtrées des ITVs*

Si on étudie l'augmentation des métriques par rapport aux travaux de 2020-2021 ou au modèle sans le rajout des matériaux des ITV, on obtient les résultats dans le tableau 12 ci-dessous. Il a également été utilisé toutes les informations des ITV sans les filtrer selon leur qualité pour justifier leur filtrage.

| Augmentions en % par rapport aux travaux 2020-2021 | | | |
|---|---|---|---|
| | Précision | Rappel | F-score |
| Modèle | 49,76 | 51,43 | 51,53 |
| Modèle + ITV | 53,93 | 57,32 | 56,99 |
| Modèle + ITV filtrées | 55,00 | 58,37 | 58,06 |

*Tableau 12 Synthétisation des améliorations réalisées*

Ainsi, la précision et le rappel sur les classes de matériaux est finalement de 0,89 et 0,87. L'ajout de ces données dans l'estimation des années de poses à fait passer le RMSE de 5,5 à 5,4 ans.

Si une ITV se passe correctement, le sens à privilégier pour une inspection est le sens d'écoulement du collecteur. Ainsi, lorsqu'un collecteur a subi plusieurs ITVs dans le même sens, seul la plus récente a été gardée. Mais s'il existe deux inspections dans le sens contraire, les deux ont été gardés puisqu'il y a eu un problème. En supprimant les doublons en faisant attention au sens d'écoulement, il a au final été gardé 17 732 tronçons avec 270 966 observations. Parmi ces observations, 23 170 amorces de branchement ont été recensées (Figure 42).

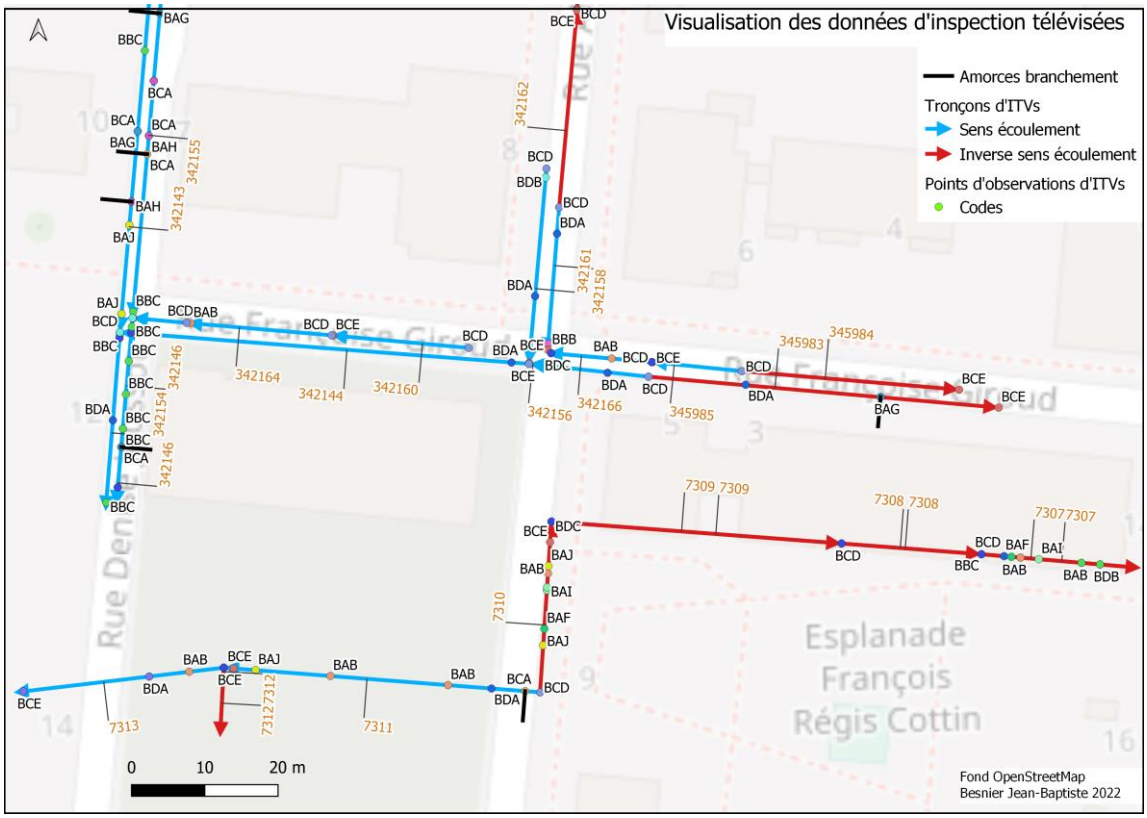



*Figure 42 Visualisation des trois couches de données  contenant les informations des ITVs*

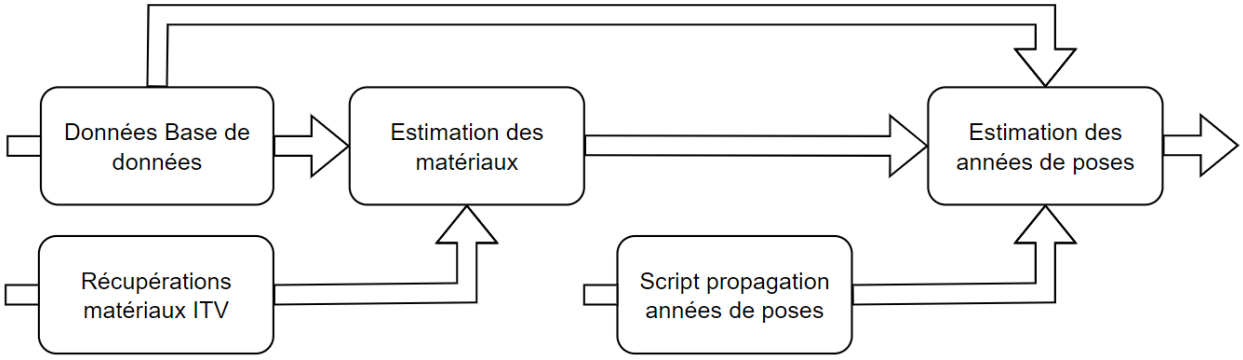



*Figure 43 Processus d'estimation complet des années de poses et des matériaux*

# Conclusion

Au cours de ces travaux, les objectifs d'amélioration de la connaissance des années de pose et de matériaux ont été remplis. À partir d'algorithmes d'apprentissage machine, la totalité des attributs manquants a été comblée. Il y a eu une progression de la connaissance et de la fiabilisation des attributs des résultats concernant les méthodes utilisées. La métropole de Lyon qui souhaitait améliorer sa connaissance du réseau possède maintenant des outils de prédiction et d'estimation puissants. Bien que la méthode n'ait pas pu être appliquée aux codes d'état de santé Indigau, les années de pose pourront intervenir dans d'autres modèles de prédiction et aider à la prévision de travaux. La remontée et l'analyse multi-source d'informations dans la base de données ont permis non seulement d'améliorer les prédictions, mais de valoriser des travaux et des données inutilisées. Le fait que la construction des lotissements ne corresponde pas à la création d'un réseau d'assainissement est un résultat en soit. La remontée des matériaux dans les fichiers des inspections télévisées puis la remontée de toutes les informations de celles-ci vont permettre aux équipes d'avoir une meilleure connaissance du terrain avec une centralisation de l'information. Au niveau des indicateurs SISPEA, ces travaux ont permis de gagner 25 points et de dépasser les attentes pour le contrat en cours. Sans le projet Hireau et ces travaux, la métropole n'aurait pas eu accès aux subventions pour entretenir et améliorer le réseau.

Un résultat un peu à part des missions du stage a représenté un vrai gain pour la métropole. L'analyse poussée et les traitements réguliers sur la couche des collecteurs a permis de mettre en lumière plus d'une centaine d'anomalies dans le SIG. Ce stage a donc aussi participé à l'amélioration de la connaissance du réseau de manière générale.

### Perspectives

Un article est en cours de rédaction pour la revue Techniques Sciences Méthode. Sa publication montrera à l'ensemble des collectivités ce qu'il est possible de mettre en œuvre au niveau informatique et géomatique pour améliorer la connaissance patrimoniale d'un réseau.

Au niveau des perspectives et des améliorations qui pourront être réalisées dans le futur, plusieurs choses peuvent être envisagées. Au niveau des regroupements des matériaux en sept classes, il faudrait envisager des relations d'inclusions entres les valeurs de matériaux. Par exemple, un collecteur est premièrement en Béton puis on viendrait spécifier s'il est armé ou non. Cela permettrait de remonter plus d'informations terrains qui n'ont pas toujours les éléments pour statuer exactement le type de matériaux. Cette méthode n'a pas été mise en place pour ne pas modifier la structure de la base en production. Un autre axe de recherche serait de réussir à qualifier la donnée produite par l'algorithme de propagation sous contrainte. I.e. enlever des années de pose connues et vérifier si l'algorithme arrive à les retrouver. Cette étape n'a pas été mise en place faute de temps. De plus, cet algorithme était venu de l'idée que les travaux se faisaient dans une même rue d'où l'étude avec les angles. D'autres idées pourraient venir améliorer les observations.

Il y aussi du travail à faire dans l'intégration, non pas des résultats, mais dans tous les processus mis en place. La métropole dispose d'un serveur de traitement FME et il y aurait lieu d'intégrer tous les traitements à un processus annuel automatique par exemple.

Enfin, une veille scientifique est à poursuivre. Au moment de l'écriture de ce rapport, une version majeure de l'outil Optuna pour optimiser les hyperparamètres est sortie implémentant de nouvelles méthodes. Il y a donc lieu de continuer à rechercher à toujours mieux valoriser la donnée.

**Formation**

Ce stage a été très valorisant dans le cadre de ma formation d'ingénieur. Il a fallu faire preuve d'autonomie et être force de proposition pour justifier des méthodes et des résultats à une équipe de manière régulière. J'ai pu acquérir des compétences dans le domaine de l'eau et des réseaux, domaines qui m'avaient attirés lors de ma formation. J'ai également beaucoup travaillé sur la vulgarisation scientifique pour faire comprendre des concepts et des problématiques complexes à des personnes de tout horizon. J'ai également réalisé une formation de descente en réseau et visité de nombreux équipements de la métropole permettant un autre regard sur le travail accompli et sur sa restitution (**annexe G** p. 82). Au niveau professionnel, j'ai pu acquérir des connaissances au sein d'une collectivité territoriale contente de mon travail fourni qui m'a proposé un poste.

# Bibliographie

**[1]** Cherqui F. et al. « Comment reconstituer l'histoire des réseaux d'assainissement et d'eau potable - Guide opérationnel - version finale ». In : *https://hireau.wordpress.com/* [En ligne]. [s.l.] : [s.n.], 2021. Disponible sur : < https://hireau.files.wordpress.com/2019/10/hireau_guide-operationnel-version-finale-3.pdf > (consulté le 4 avril 2022)

**[2]** Direction de l'eau. *Schéma Général d'assainissement Du Grand Lyon*. [s.l.] : [s.n.], 2015.

**[3]** Niogret P., Chuzeville G. *APPROCHE GÉOMATIQUE POUR LA RECONSTITUTION DE L'ANNÉE DE POSE DU RÉSEAU D'ASSAINISSEMENT Par Pierre NIOGRET*. [En ligne]. [s.l.] : [s.n.], 2020. Disponible sur : < https://hireau.files.wordpress.com/2021/01/niogret_pierre_rapport_stage_m2.pdf >

**[4]** Le Moyec T., Chuzeville G. *L'amélioration De La Connaissance Du Patrimoine Entreprise : Métropole De Lyon*. [En ligne]. [s.l.] : [s.n.], 2021. Disponible sur : < https://hireau.files.wordpress.com/2021/10/le-moyec_tristan_memoire.pdf >

**[5]** Shalev-Shwartz S. *IFT-7002 Fondements De l'apprentissage Machine*. [En ligne]. [s.l.] : [s.n.], 2021. Disponible sur : < http://www2.ift.ulaval.ca/~mmarchand/IFT7002/ModeleGeneral.pdf >

**[6]** Kontorovich A., Pinelis I. *Exact Lower Bounds for the Agnostic Probably-Approximately-Correct (PAC) Machine Learning Model*. [En ligne]. [s.l.] : [s.n.], 2017. Disponible sur : < https://arxiv.org/pdf/1606.08920.pdf >

**[7]** J2kun. « Probably Approximately Correct — a Formal Theory of Learning ». In : *Math ∩ Programming* [En ligne]. [s.l.] : [s.n.], 2014. Disponible sur : < https://jeremykun.com/2014/01/02/probably-approximately-correct-a-formal-theory-of-learnin/ >

**[8]** Fort G., Lerasle M., Moulines E. *Statistique Et Apprentissage Notes Du Cours MAP433*. [En ligne]. [s.l.] : [s.n.], 2020. Disponible sur : < https://lerasle.perso.math.cnrs.fr/docs/mainpoly.pdf >

**[9]** Fernández-Delgado M. et al. « Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? » *Journal of Machine Learning Research* [En ligne]. 2014. Vol. 15, p. 3133-3181. Disponible sur : < https://jmlr.csail.mit.edu/papers/volume15/delgado14a/delgado14a.pdf >

**[10]** Olson R. S. et al. « Data-driven Advice for Applying Machine Learning to Bioinformatics Problems ». *arXiv:1708.05070 [cs, q-bio, stat]* [En ligne]. 7 janvier 2018. Disponible sur : < https://arxiv.org/abs/1708.05070 >

**[11]** « XGBoost Documentation — Xgboost 1.5.0-dev Documentation ». In : *xgboost.readthedocs.io* [En ligne]. [s.l.] : [s.n.], 2014. Disponible sur : < https://xgboost.readthedocs.io/en/latest/index.html >

**[12]** Hachcham A. « XGBoost: Everything You Need to Know ». In : *neptune.ai* [En ligne]. [s.l.] : [s.n.], 2021. Disponible sur : < https://neptune.ai/blog/xgboost-everything-you-need-to-know >

**[13]** Laboratoire de l'Accélérateur Linéaire (LAL). « Higgs Boson Machine Learning Challenge ». In : *kaggle.com* [En ligne]. [s.l.] : [s.n.], 2014. Disponible sur : < https://www.kaggle.com/competitions/higgs-boson/data >

**[14]** Stéphanie G. « Les Différences Entre Arbre De décision, Random Forest Et Gradient Boosting ». In : *LeMagIT* [En ligne]. [s.l.] : [s.n.], 2021. Disponible sur : < https://www.lemagit.fr/conseil/Les-differences-entre-arbre-de-decision-Random-Forest-et-Gradient-Boosting >

**[15]** Starmer J. *XGBoost* [En ligne]. *YouTube*. 2019. Disponible sur : < https://www.youtube.com/watch?v=OtD8wVaFm6E&list=PLblh5JKOoLULU0irPgs1SnKO6wqVjKUsQ >

**[16]** Baudoux L., Inglada J., Mallet C. « Toward a Yearly Country-Scale CORINE Land-Cover Map without Using Images: a Map Translation Approach ». *Remote Sensing* [En ligne]. 11 mars 2021. Vol. 13, n°6, p. 1060. Disponible sur : < https://doi.org/10.3390/rs13061060 > (consulté le 13 avril 2022)

**[17]** Shahul ES. « 7 Cross-Validation Mistakes That Can Cost You a Lot [Best Practices in ML] ». In : *neptune.ai* [En ligne]. [s.l.] : [s.n.], 2021. Disponible sur : < https://neptune.ai/blog/cross-validation-mistakes >

**[18]** « Optuna: A hyperparameter optimization framework — Optuna 2.10.0 documentation ». In : optuna.readthedocs.io [En ligne]. [s.l.] : [s.n.], 2018. Disponible sur : < https://optuna.readthedocs.io/en/stable/index.html >

**[19]** Imamura H. « Benchmarks with Kurobako · optuna/optuna Wiki ». In : GitHub [En ligne]. [s.l.] : [s.n.], 2020. Disponible sur : < https://github.com/optuna/optuna/wiki/Benchmarks-with-Kurobako >

**[20]** Wang C., Deng C., Wang S. *Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary label-imbalanced Classification with XGBoost a Preprint*. [En ligne]. [s.l.] : [s.n.], 2021. Disponible sur : < https://arxiv.org/pdf/1908.01672.pdf >

**[21]** Chawla N. V. et al. « SMOTE: Synthetic Minority Over-sampling Technique ». *Journal of Artificial Intelligence Research* [En ligne]. 1 juin 2002. Vol. 16, n°16, p. 321-357. Disponible sur : < https://doi.org/10.1613/jair.953 >

**[22]** Northcutt C. G. « Cleanlab: the History, Present, and Future ». In : Cleanlab [En ligne]. [s.l.] : [s.n.], 2022. Disponible sur : < https://cleanlab.ai/blog/cleanlab-history/ >

**[23]** Northcutt C. G., Jiang L., Chuang I. L. « Confident Learning: Estimating Uncertainty in Dataset Labels *». arXiv:1911.00068 [cs, stat]* [En ligne]. 21 août 2022. Disponible sur : < https://arxiv.org/abs/1911.00068 >

**[24]** Kassambara A. « Kappa De Cohen Dans R». In : *Datanovia* [En ligne]. [s.l.] : [s.n.], [s.d.]. Disponible sur : < https://www.datanovia.com/en/fr/lessons/kappa-de-cohen-dans-r-pour-deux-variables-categorielles/ >

# Table des illustrations

## Table des tableaux



## Annexes

### Sommaire

## Annexe A : Justification apprentissage machine

Supposons que chaque $x_i$ est indépendamment échantillonné selon une distribution $D$ inconnue. i.e $D$ est une distribution sur le domaine $X$. Supposons également que $Y$ est déterminé par $f \in H, H \subset Y^X$ et qu'il existe une relation entre $S$ et $(D, f)$. L'objectif est de trouver $h : h \approx f$. On note alors $L_{D,f}(h)$ l'erreur de généralisation de $h$ par $f$ dans le cadre $D$.

$$Erreur\ de\ généralisation : \ L_{D,f}(h) = P_{X\sim D}[h(x) \neq f(x)] = D(\{x \in X : h(x) \neq f(x)\})$$

Cependant, l'algorithme ne connait ni la distribution $D$ ni la fonction $f$ déterminant $Y$. On fournit alors à l'algorithme un paramètre de précision $\epsilon$ ainsi qu'un paramètre de confiance $\delta$. Le but est alors de trouver en utilisant $S$, un apprenant $A$ donnant l'hypothèse $A(S)$ tel que :

$$\textit{Critère PAC :} \qquad A : \ P(L_{D,f}(A(S)) \leq \epsilon) \ \geq 1 - \delta$$

Autrement dit, l'algorithme avec le modèle A doit être Probablement ($\geq 1 - \delta$) Approximativement ($\leq \epsilon$) Correct. Cette précision $\epsilon$ ne doit pas donc pas être atteinte uniformément mais avec la probabilité $1 - \delta$.

Comme $L_{D,f}(h)$ ne peut être connu, pour contourner ce problème, on utilise la Minimisation du Risque Empirique (ERM) qui pour $S$ et $h$ donné s'écrit :

$$\textit{Risque empirique :} \qquad L_S(h) = \frac{1}{m}|\{i : h(x_i) \neq y_i\}|$$

$$\textit{Minimisation du Risque Empirique :} \qquad ERM_{H(S)} = \min_{h \in H} L_S(h)$$

<u>Théorème 1:</u>

Soit $H$ une classe d'hypothèses finie.

$$\forall (\varepsilon, \delta) \in R^{*+^2}, \forall\, f \in H, H \subset Y^X, \forall\, S \in (X \times Y)^n, X\sim D$$

$$n \geq \frac{log\left(\frac{|H|}{\delta}\right)}{\varepsilon} \Rightarrow \ P(L_{D,f}(ERM_{H(S)}) \leq \varepsilon) \geq 1 - \delta$$

Ainsi, en étudiant simplement le nombre d'échantillons disponible n, en connaissant le nombre de classes à prédire et en connaissant la dimension des vecteurs de données m, nous pouvons savoir si notre jeu est apprenable au sens PAC. C'est-à-dire qu'il existe un modèle $h$ tel que pour n'importe quelle distribution inconnue $D$, $h(S)$ est probablement approximativement correct.

Cependant, cette estimation au sens PAC présente plusieurs défauts. Le premier étant qu'on impose une hypothèse extrêmement forte sur l'existence d'une fonction $f$ amenant à une erreur de généralisation nulle. Un moyen de se rendre compte de la contrainte de cette hypothèse est qu'il est impossible d'avoir des données bruitées dans $S$. Sinon f ne peut pas avoir une erreur de généralisation nulle car elle fera forcément des erreurs sur le jeu de test où les $y_i$ ne sont pas connus. Si on prend l'exemple de l'estimation des matériaux, il est possible qu'une canalisation en béton parmi deux identiques voisines ait été remplacée par une autre en plastique. La seule différence entre ces deux canalisations presque identiques est alors le type de matériaux. Il ne peut donc pas exister une fonction de prédiction f déterminant parfaitement le jeu de donnée.

Pour pallier cela, il faut se pencher sur le PAC agnostique. On supposera maintenant que $Y$ suit également une distribution inconnue. On a maintenant $X \times Y \sim D$ où D est la distribution jointe sur $X \times Y$.

On redéfini donc :

$$Erreur\ de\ généralisation :\ L_D(h) = P_{(x,y)\sim D}[h(x) \neq y] = D(\{(x,y) : h(x) \neq y\})$$

$$Critère\ PAC :\qquad A :\ P(L_D(A(S)) \leq \min_{h\in H} L_D(h) + \epsilon)\ \geq 1-\delta$$

Théorème 2:

Soit $H$ une classe finie.

$$\forall(\varepsilon,\delta) \in R^{*+^2}, \forall\, f \in H, H \subset Y^X, \forall\, S \in (X\times Y)^n, X\sim D$$

$$n \geq \frac{2\,log\left(\frac{2|H|}{\delta}\right)}{\varepsilon^2} \Rightarrow\ P(L_D\left(ERM_{H(S)}\right) \leq \min_{h\in H} L_D(h) + \epsilon)\ \geq 1-\delta$$

## Annexe B : Construction fonction objectif et fonctionnement XGBoost

Dans la suite de l'explication du XGBoost, par soucis de simplicité, sera utilisée la fonction de perte quadratique.

$$L(\theta) = \sum_i (y_i - \hat{y}_i)^2$$

Comme présenté, l'algorithme construit sa prédiction à partir de « weak learners ». Contrairement à un arbre de décision classique, l'algorithme va construire des CART (Classification And Regression Trees) où à chaque feuille sera associé un score permettant de témoigner de l'importance de cette feuille dans la classification. La prédiction $\hat{y}_i$ est alors définit comme :

$$\hat{y}_i = \sum_{k=1}^{T} f_k(x_i), \begin{cases} T \in \mathbb{N}, nombre\ d'arbres \\ f_k \in \mathcal{F} \\ \mathcal{F}\ ensemble\ des\ fonctions\ définissant\ des\ CARTs \end{cases}$$

À partir de cette estimation, il est possible de redéfinir la fonction objectif (1) qui devient alors :

$$obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{T} w(f_k) \qquad (2)$$

Où $w(f_k)$ est un terme traduisant la complexité de l'arbre k permettant la régularisation.

L'entrainement se fait par ajout successif de CARTs qui apprennent des erreurs des précédents. Mathématiquement, nous pouvons modifier (2) pour transcrire l'entrainement.

$$\widehat{y_i}^{(0)} = 0$$

$$\widehat{y_i}^{(1)} = f_1(x_i) = \widehat{y_i}^{(0)} + f_1(x_i)$$

$$\widehat{y_i}^{(2)} = f_1(x_i) + f_2(x_i) = \widehat{y_i}^{(1)} + f_2(x_i)$$

…

$$\widehat{y_i}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \widehat{y_i}^{(t-1)} + f_t(x_i)$$

D'où pour l'arbre t :

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}^{(t)}) + \sum_{i=1}^{t} w(f_i) \qquad (3)$$

En utilisant la formule de la perte quadratique on obtient :

$$= \sum_{i=1}^{n} (y_i - (\widehat{y_i}^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^{t} w(f_i)$$

$$= \sum_{i=1}^{n} [2(\widehat{y_i}^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + w(f_t) + C, C \in \mathbb{R}$$

Dans C est contenu tous les éléments ne dépendant pas de t. En utilisant le développement de Taylor à l'ordre 1 sur $obj^{(t)}$ , on obtient :

$$obj^{(t)} = \sum_{i=1}^{n} [l(y_i, \widehat{y_i}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i)] + w(f_t) + C', C' \in \mathbb{R}$$

Avec :

$$g_i = \frac{\partial l(y_i, \widehat{y_i}^{(t-1)})}{\partial \widehat{y_i}^{(t-1)}} \; et \; h_i = \frac{\partial^2 l(y_i, \widehat{y_i}^{(t-1)})}{\partial^2 \widehat{y_i}^{(t-1)}} \; Gradient \; et \; Hessienne \; de \; l(y_i, \widehat{y_i}^{(t-1)})$$

Finalement, en retirant les éléments ne dépendant pas de la construction de l'arbre t on obtient :

$$obj^{(t)} = \sum_{i=1}^{n} [\, g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i)] + w(f_t) \qquad (4)$$

Il faut noter ici, que même si le développement de Taylor doit se faire à l'ordre 2 pour la fonction de perte *logloss*, on retrouvera une fonction objectif de la forme de (4). La force de cet algorithme réside dans le fait qu'il n'a besoin que de $g_i$ et $h_i$ pour optimiser la fonction de perte. Il reste maintenant à expliciter le terme de régularisation $w(f_t)$.

Pour cela on pose $f_t(x) = w_{q(x)}, w \in \mathbb{R}^T, q : \mathbb{R}^D \to \{1,2, \dots, T\}$ . $w$ est le vecteur des scores de chaque feuille. $q$ est la fonction qui associe à chaque donnée sa feuille sur un arbre de T feuilles. On peut alors construire $w(f)$.

$$w(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {w_j}^2, (\gamma, \lambda) \in \mathbb{R}^2$$

Et remplacer $w(f)$ dans (4) :

$$obj^{(t)} = \sum_{i=1}^{n} [\, g_i w_{q(x_i)} + \frac{1}{2} h_i {w_{q(x_i)}}^2 (x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {w_j}^2$$

$$obj^{(t)} = \sum_{j=1}^{T} [\, (\sum_{i \in I_j} g_i)\, w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda)\, {w_j}^2] \; + \gamma T \,, I_j = \{i | q(x_i) = j\} \qquad (5)$$

$I_j$ étant l'ensemble des indices des données assignées à la feuille j, le changement d'indice de sommation et le regroupement par les $w_j$ est possible car toutes les données ont le même score sur une même feuille.

On pose ensuite

$$G_j = \sum_{i \in I_j} g_i \;\; et \; H_j = \sum_{i \in I_j} h_i$$

D'où (5) devient :

$$obj^{(t)} = \sum_{j=1}^{T} [\, G_j w_j + \frac{1}{2}(H_j + \lambda) {w_j}^2] \; + \gamma T \qquad (6)$$

Dans (6), la meilleure réduction possible de $G_j w_j + \frac{1}{2}(H_j + \lambda) {w_j}^2$ est :

$${w_j}^* = \frac{-G_j}{H_j + \lambda}$$

D'où

$$obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \qquad (7)$$

Cette dernière équation (7) est donc la fonction objectif qui permet de mesurer la qualité de la structure d'un arbre $q(x)$.

# Annexe C : Fiche P202.2B SISPEA – Partie B

## Partie B : Inventaire des réseaux (30 points)

- **10 points (VP.252, VP.253 et VP.254) - les 10 points sont acquis si les 2 conditions suivantes sont remplies :**
  - **Existence d'un inventaire des réseaux** identifiant les tronçons de réseaux avec mention du linéaire de la canalisation, de la catégorie de l'ouvrage définie en application de l'article R. 554-2 du code de l'environnement ainsi que de la précision des informations cartographiques définie en application du V de l'article R. 554-23 du même code (VP.252) et, **pour au moins la moitié du linéaire total des réseaux**, les informations sur les matériaux et les diamètres des canalisations de collecte et de transport des eaux usées (VP.253)
  - **La procédure de mise à jour** du plan des réseaux est complétée en y intégrant la mise à jour de l'inventaire des réseaux (VP.254)
- **De 1 à 5 points (VP.253) :** Lorsque les informations sur les matériaux et les diamètres sont rassemblées pour la moitié du linéaire total des réseaux, **un point supplémentaire est attribué chaque fois que sont renseignés 10% supplémentaires du linéaire total, jusqu'à 90%.** Le cinquième point est accordé lorsque les informations sur les matériaux et les diamètres sont rassemblées pour au moins 95% du linéaire total des réseaux :

  Matériaux et diamètres connus pour 60% à 69,9% du linéaire des réseaux : 1 point supplémentaire

  Matériaux et diamètres connus pour 70% à 79,9% du linéaire des réseaux : 2 points supplémentaires

  Matériaux et diamètres connus pour 80% à 89,9% du linéaire des réseaux : 3 points supplémentaires

  Matériaux et diamètres connus pour 90% à 94,9% du linéaire des réseaux : 4 points supplémentaires

  Matériaux et diamètres connus pour au moins 95% du linéaire des réseaux : 5 points supplémentaires

**Si la procédure de mise à jour du plan des réseaux n'est pas complétée de la mise à jour de l'inventaire des réseaux (VP.254), les points de la VP.252 et de la VP.253 ne sont pas comptabilisés.**

- **De 0 à 15 points (VP.255) - indépendante des VP.252, VP.253 et VP.254 :**

  **L'inventaire des réseaux mentionne pour chaque tronçon la date ou la période de pose** des tronçons identifiés à partir du plan des réseaux, la moitié (50%) du linéaire total des réseaux étant renseigné. Lorsque les informations sur les dates ou périodes de pose sont rassemblées pour la moitié du linéaire total des réseaux, **un point supplémentaire est attribué chaque fois que sont renseignés 10% supplémentaires du linéaire total, jusqu'à 90%.** Le cinquième point est accordé lorsque les informations sur les dates ou périodes de pose sont rassemblées pour au moins 95% du linéaire total des réseaux

  Dates ou périodes de pose connues pour moins de 49,9% du linéaire des réseaux : 0 point

  Dates ou périodes de pose connues pour ≥ 50% à 59,9% du linéaire des réseaux : 10 points (Cas 0)

  Dates ou périodes de pose connues pour ≥ 60% à 69,9% du linéaire des réseaux : 11 points (Cas 1)

  Dates ou périodes de pose connues pour ≥ 70% à 79,9% du linéaire des réseaux : 12 points (Cas 2)

  Dates ou périodes de pose connues pour ≥ 80% à 89,9% du linéaire des réseaux : 13 points (Cas 3)

  Dates ou périodes de pose connues pour ≥ 90% à 94,9% du linéaire des réseaux : 14 points (Cas 4)

  Dates ou périodes de pose connues pour ≥ 95% du linéaire des réseaux : 15 points (Cas 5)

## Annexe D : Tableaux

Le tableau suivant mesure l'accord sur les matériaux entre les données de la base de données (vertical) et des données reçues des ITV (horizontal). En calculant l'indice de Kappa pour deux variables catégorielles à niveaux multiples, un accord de 0.35 correspond à un accord faible

| Effectifs | AMCI | AUTR | BTAM | BTAU | FON | PLAS | ROCH | Effectifs marginaux |
|---|---|---|---|---|---|---|---|---|
| AMCI | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| AUTR | 0 | 0 | 5 | 1 | 0 | 19 | 0 | 25 |
| BTAM | 22 | 89 | 3387 | 217 | 45 | 143 | 2 | 3905 |
| BTAU | 16 | 67 | 1119 | 126 | 0 | 109 | 3 | 1440 |
| FON | 0 | 2 | 20 | 0 | 3 | 9 | 0 | 34 |
| PLAS | 0 | 60 | 334 | 31 | 0 | 861 | 0 | 1286 |
| ROCH | 0 | 2 | 51 | 4 | 0 | 12 | 21 | 90 |
| Marge des colonnes | 43 | 220 | 4916 | 379 | 48 | 1153 | 26 | 6785 |
| Proportions marginales | AMCI | AUTR | BTAM | BTAU | FON | PLAS | ROCH | Effectifs marginaux |
| AMCI | 7,37E-04 | 0,00E+00 | 0,00E+00 | 0,00E+00 | 0,00E+00 | 0,00E+00 | 0,00E+00 | 7,37E-04 |
| AUTR | 0,00E+00 | 0,00E+00 | 7,37E-04 | 1,47E-04 | 0,00E+00 | 2,80E-03 | 0,00E+00 | 3,68E-03 |
| BTAM | 3,24E-03 | 1,31E-02 | 4,99E-01 | 3,20E-02 | 6,63E-03 | 2,11E-02 | 2,95E-04 | 5,76E-01 |
| BTAU | 2,36E-03 | 9,87E-03 | 1,65E-01 | 1,86E-02 | 0,00E+00 | 1,61E-02 | 4,42E-04 | 2,12E-01 |
| FON | 0,00E+00 | 2,95E-04 | 2,95E-03 | 0,00E+00 | 4,42E-04 | 1,33E-03 | 0,00E+00 | 5,01E-03 |
| PLAS | 0,00E+00 | 8,84E-03 | 4,92E-02 | 4,57E-03 | 0,00E+00 | 1,27E-01 | 0,00E+00 | 1,90E-01 |
| ROCH | 0,00E+00 | 2,95E-04 | 7,52E-03 | 5,90E-04 | 0,00E+00 | 1,77E-03 | 3,10E-03 | 1,33E-02 |
| Marge des colonnes | 6,34E-03 | 3,24E-02 | 7,25E-01 | 5,59E-02 | 7,07E-03 | 1,70E-01 | 3,83E-03 | 1,00E+00 |

| | |
|---|---|
| proportion de concordance observée | 0,64893147 |
| proportion d'un accord aléatoire | 0,46127121 |
| Kappa | 0,34833902 |

| | |
|---|---|
| Intervalle de confiance Kappa | 5,94928E-05 |
| intervalle de confiance à 95% | 0,000116606 |

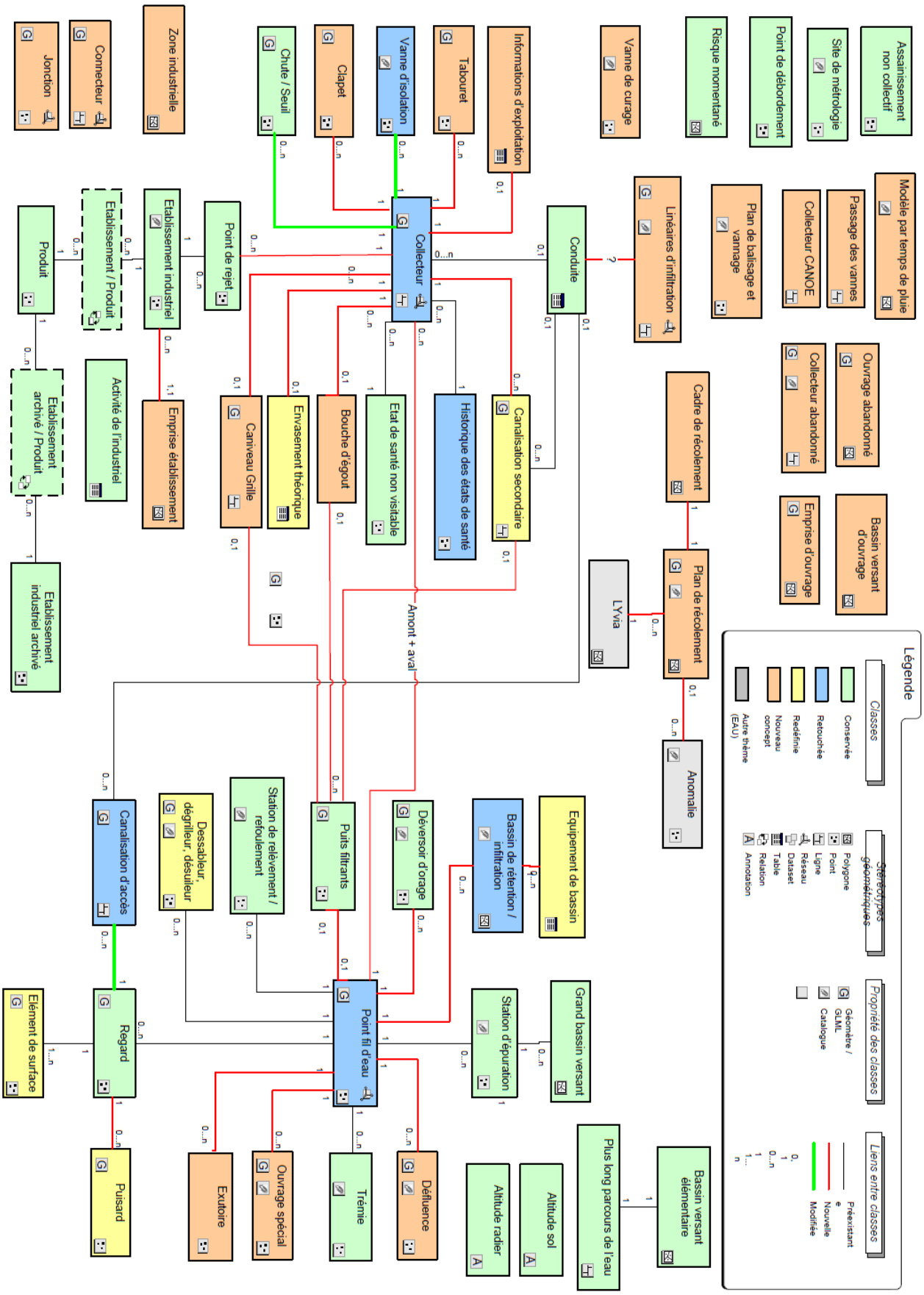
Le tableau suivant mesure la précision de la classe majoritaire venant des archives des inspections des collecteurs visitables. Seuls les codes BETONBETON et MACONNERBE ont été remonté.

| | ROCH | PLAS | FON | BTAU | BTAM | AUTR | AMCI | Précision classe majoritaire |
|---|---|---|---|---|---|---|---|---|
| MACONNERMA | 13 | 0 | 0 | 0 | 21 | 0 | 0 | 0,61764706 |
| Maçonnerie | 483 | 1 | 0 | 35 | 259 | 14 | 3 | 0,60754717 |
| MACONNEREN | 89 | 7 | 0 | 806 | 498 | 34 | 10 | 0,55817175 |
| ENDUITENDU | 2 | 0 | 0 | 13 | 8 | 0 | 0 | 0,56521739 |
| Enduit | 392 | 4 | 0 | 535 | 652 | 54 | 0 | 0,39828955 |
| BETONMACON | 24 | 0 | 0 | 34 | 42 | 0 | 0 | 0,42 |
| BETONENDUI | 21 | 3 | 0 | 153 | 118 | 11 | 0 | 0,5 |
| BETONBETON | 2 | 0 | 0 | 19 | 97 | 0 | 1 | 0,81512605 |
| Béton | 159 | 30 | 1 | 1773 | 2356 | 60 | 28 | 0,53460404 |
| MACONNERBE | 0 | 0 | 0 | 5 | 12 | 0 | 0 | 0,70588235 |
| ENDUITMACO | 0 | 0 | 0 | 10 | 10 | 0 | 0 | 0,5 |
| ENDUITBETO | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0,5 |

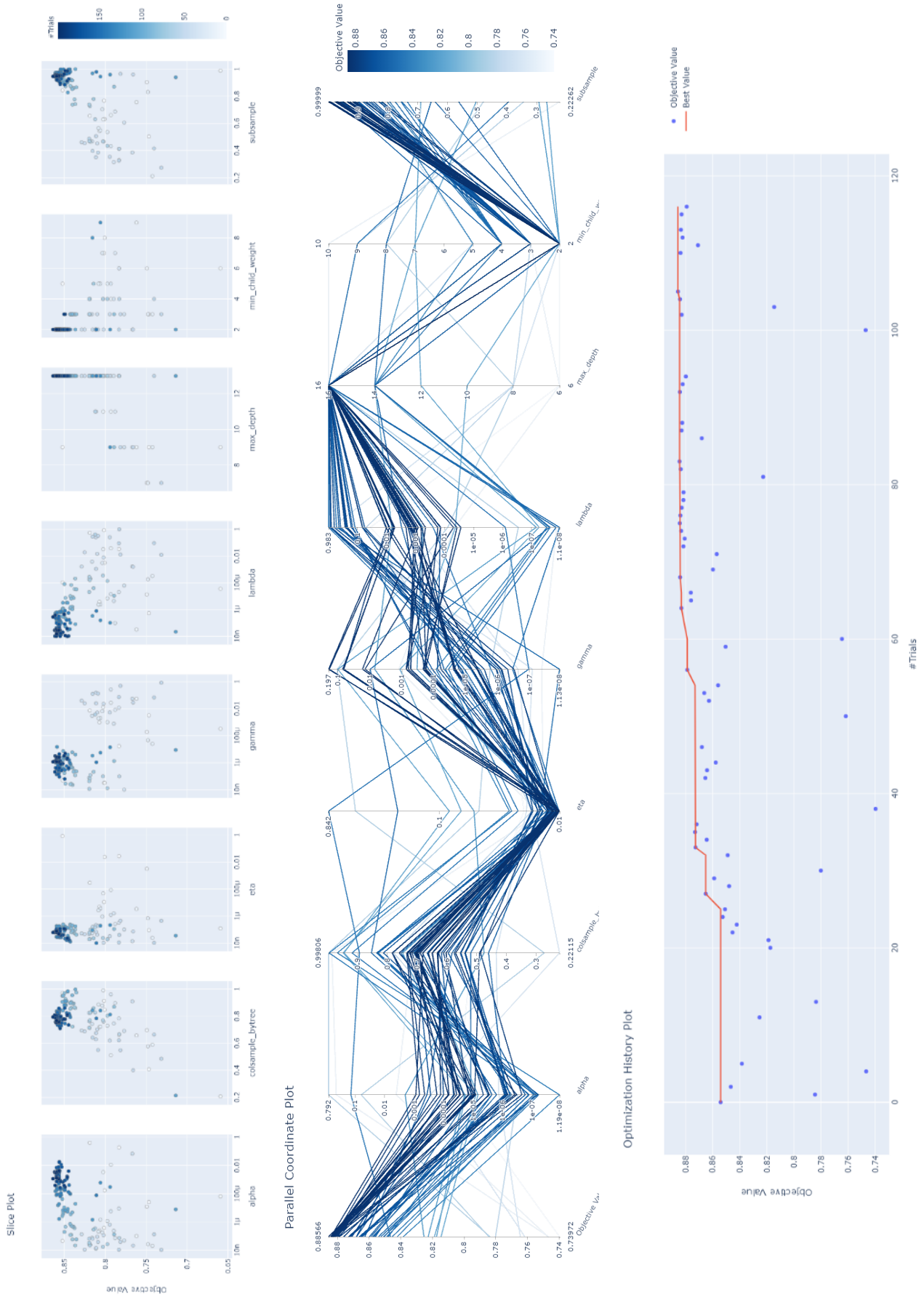
Tableau d'appariement entre les données de la base de données et les informations des matériaux remontées par les Inspections TéléVisées. Seuls les codes AMCI, BTAM et PLAS ont été remontés car ils avaient une précision assez élevée. Le code AUTR de la base de donnée a été remplacé quand de l'information était disponible puisque c'est une classe englobante.

| | | AMCI | AUTR | BTAM | BTAU | FON | PLAS | ROCH | Somme | Précision |
|---|---|---|---|---|---|---|---|---|---|---|
| ITV | AMCI | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 1 |
| | AUTR | 0 | 0 | 5 | 1 | 0 | 19 | 0 | 25 | 0 |
| | BTAM | 22 | 89 | 3387 | 217 | 45 | 143 | 2 | 3905 | 0,86734955 |
| | BTAU | 16 | 67 | 1119 | 126 | 0 | 109 | 3 | 1440 | 0,0875 |
| | FON | 0 | 2 | 20 | 0 | 3 | 9 | 0 | 34 | 0,08823529 |
| | PLAS | 0 | 60 | 334 | 31 | 0 | 861 | 0 | 1286 | 0,66951789 |
| | ROCH | 0 | 2 | 51 | 4 | 0 | 12 | 21 | 90 | 0,23333333 |
| | Somme | 43 | 220 | 4916 | 379 | 48 | 1153 | 26 | 6785 | 0,42084801 |
| | Rappel | 0,11627907 | 0 | 0,68897478 | 0,33245383 | 0,0625 | 0,74674762 | 0,80769231 | 0,39352109 | |

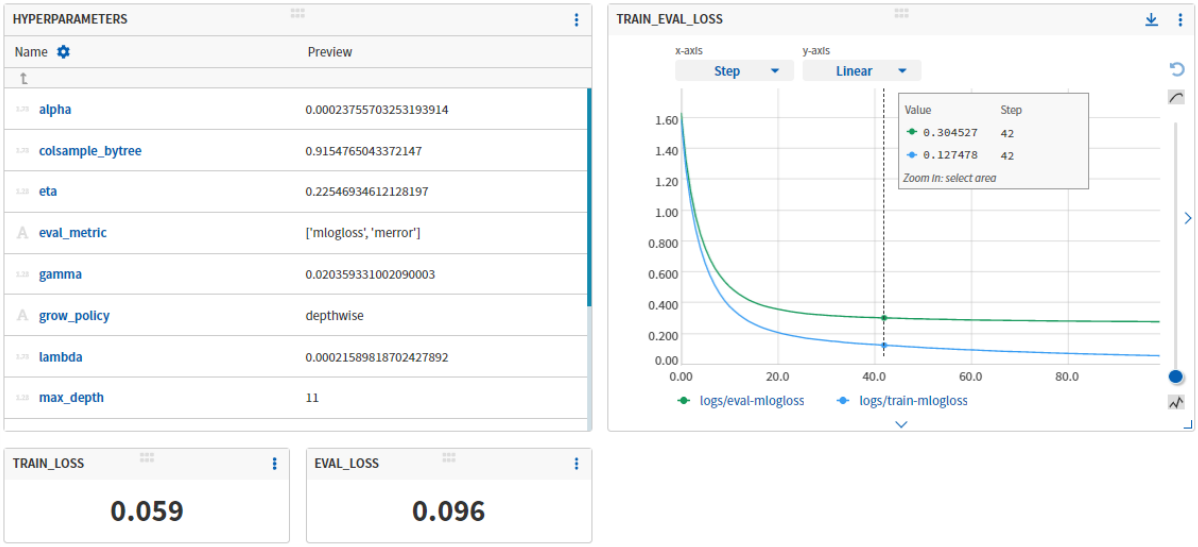
# Annexe E : Modèle conceptuel de la BDD assainissement simplifié



# Annexe F : Optimisation Hyperparamètres



Le premier et le deuxième graphique montrent au fur et à mesure des itérations, la convergence de chaque hyperparamètre vers une solution pour optimiser la fonction objectif. Le troisième graphique, mets en valeur au fur et à mesure des itérations, la meilleure batterie de paramètre.



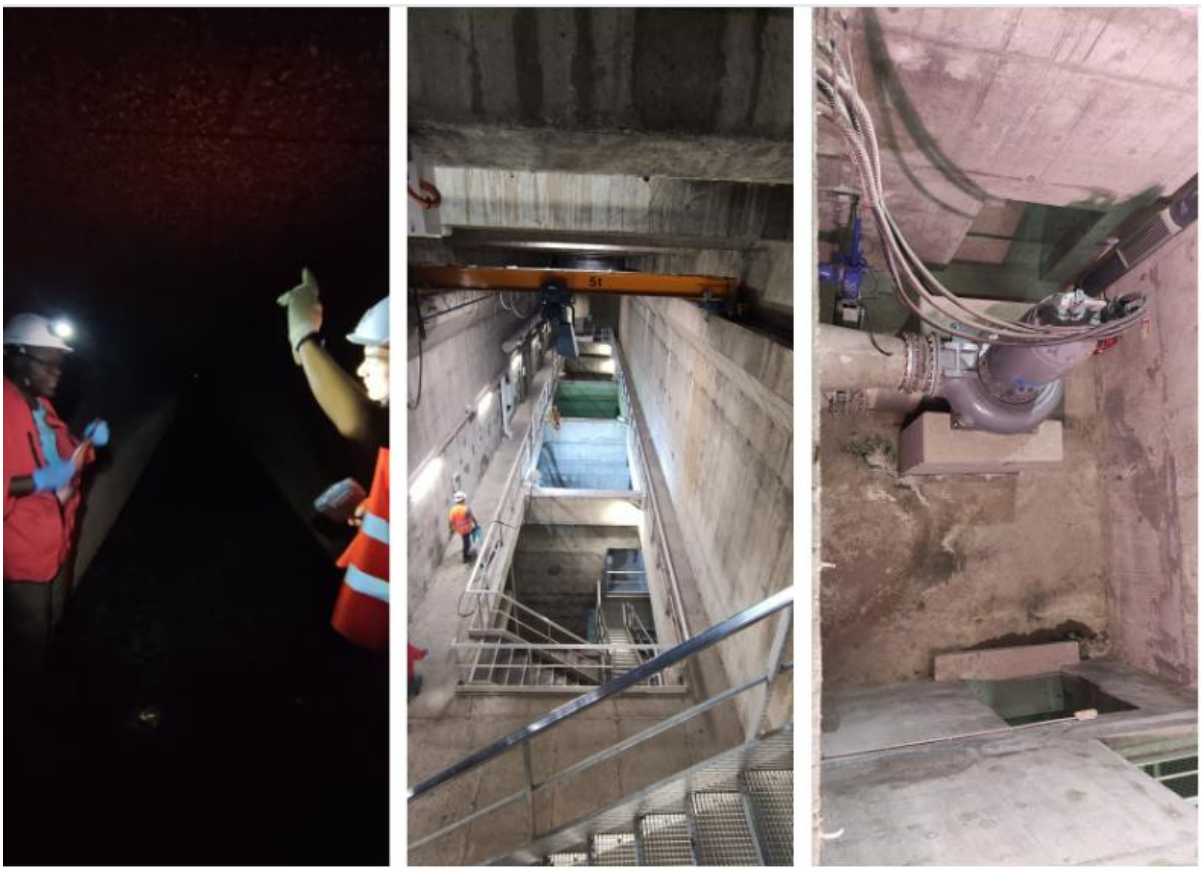

Le Tableau de bord ci-dessus, est tiré de l'outil en ligne Neptuna (https://neptune.ai/) auquel a été connecté le notebook python sur Google Colab. Cela permettait de suivre en temps réel l'évolution des métriques et surtout de sauvegarder chaque test pour les réutiliser par la suite. Cette partie n'a pas été présentée aux équipes car ce tableau sert uniquement pour le développement.

## Annexe G : Visites réalisées

Visite de travaux en bord du Rhône pour raccorder des immeubles historiques au réseau d'assainissement.

Visite de la station d'épuration de Pierre Bénite (bassin où des bactéries dégradent les éléments en suspension dans l'eau).

Mesures du débit de ruisseaux de la métropole et études d'anomalies (collecteurs bouchés)

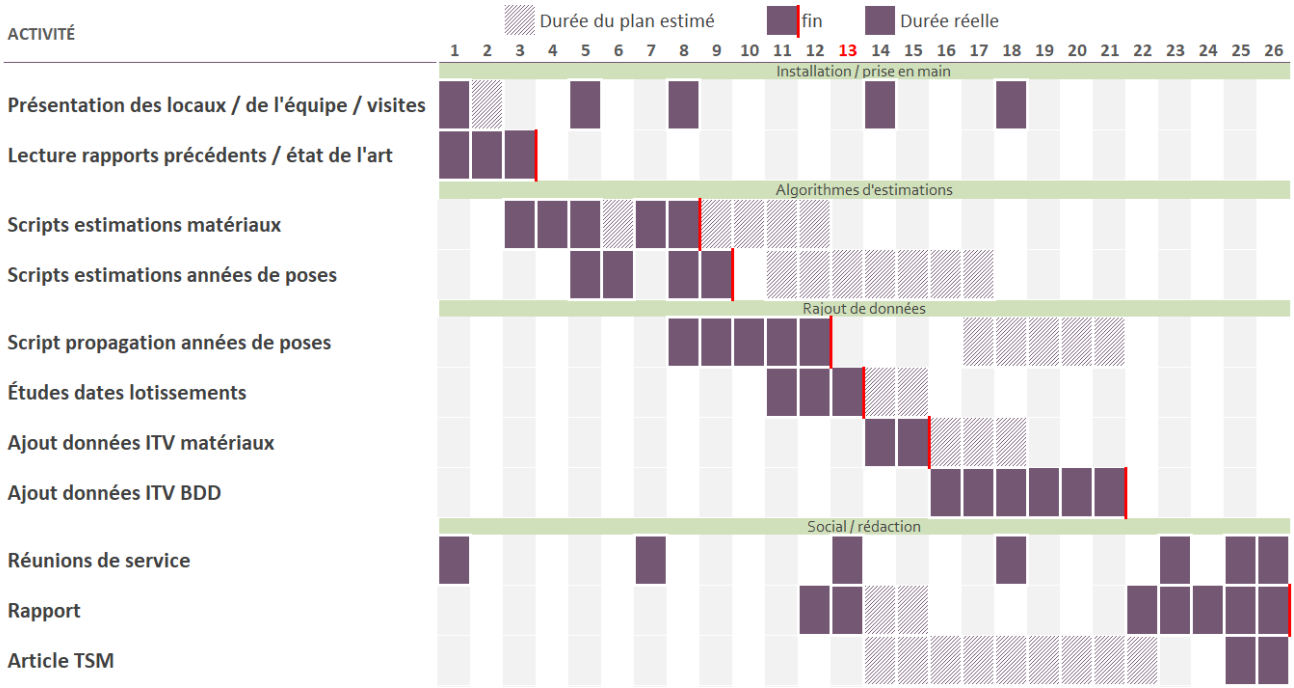


Visite du déversoir d'orage de la station de pierre bénite. C'est un collecteur rejetant directement dans le Rhône via débordement au-dessus d'une digue d'un canal. Visite également du système de remonté des eaux du réseau avant l'arrivée à la station pour qu'elles puissent ensuite s'écouler de manière gravitaire


Visite des hangars des camions hydro cureurs. Ce sont eux qui viennent curer le réseau grâce à des pompes et de l'eau sous pression. Cette visite était un moyen pour les dessinateurs de dimensionner les projets futurs pour un passage aisé des camions (d'où les plots simulant des trottoirs).

# Annexe H : Suivi du stage

Diagramme de Gantt




Outil Trello pour organiser et prioriser le travail