



HAL
open science

Gestion Patrimoniale Multi-Echelles des réseaux d'eau potable (GePaME) - Rapport intermédiaire 2022

Alain Husson, Aurélien Mirebeau, Achille Payant, Bénédicte Rulleau, Eddy Renaud, Kevin Caillaud, Marie Collet, Nicolas Rodriguez, Selma Baati, Simon Fidelle, et al.

► To cite this version:

Alain Husson, Aurélien Mirebeau, Achille Payant, Bénédicte Rulleau, Eddy Renaud, et al.. Gestion Patrimoniale Multi-Echelles des réseaux d'eau potable (GePaME) - Rapport intermédiaire 2022. INRAE UR ETTIS; SWDE. 2022. hal-03836891

HAL Id: hal-03836891

<https://hal.inrae.fr/hal-03836891v1>

Submitted on 2 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Projet GePaME

Gestion Patrimoniale Multi-Echelles des réseaux d'eau potable

Rapport intermédiaire 2022

Alain Husson, Aurélien Mirebeau, Achille Payant, Bénédicte Rulleau, Eddy Renaud, Kevin Caillaud, Marie Collet, Nicolas Rodriguez, Selma Baati, Simon Fidelle, Yves Le Gat

02/09/2022



Table des matières

1.	Introduction.....	12
1.1.	Contexte	12
1.2.	Contenu du projet GePaME.....	12
1.3.	Rappel des avancées de la première année du projet	12
1.4.	Contenu du rapport.....	13
2.	Tâche 01 – Prise en main et traitement des données.....	14
2.1.	Génération des fichiers pour Casses sur l’ensemble du territoire SWDE	14
2.1.1.	Rappel du contexte.....	14
2.1.2.	Prise en compte des types d’activités pertinents pour les pertes en eau	14
2.1.3.	Adaptations suite à l’application sur les secteurs de Mons, Ath, et Gaurain.....	15
2.1.4.	Autres adaptations de la méthode.....	16
2.2.	Validation de la méthode d’affectation des défaillances de conduites aux tronçons du SIG19	16
2.3.	Bilan de l’affectation des défaillances de conduites aux tronçons du SIG	20
2.3.1.	Résumé	20
2.3.2.	Phénomène de concentration des défaillances de conduites	21
2.4.	Étude de l’impact de la complétion de données sur les prédictions de défaillances	22
2.4.1.	Contexte et objectifs	22
2.4.2.	Résumé de l’étude.....	22
2.5.	Préparation des données issues de la RF pour la tâche 07	23
2.5.1.	Besoins en données de la tâche 07 et données disponibles	23
2.5.2.	Nettoyage des fichiers de prestations de RF.....	23
2.5.3.	Regroupement en campagnes de RF.....	24
2.5.4.	Lien entre campagnes de RF et fuites trouvées	24
2.5.5.	Table d’entrée pour la régression logistique.....	25
3.	Tâche 02 – Analyse de la gouvernance de la SWDE en matière de GPI	25
3.1.	Présentation de la tâche 02.....	25
3.2.	Appropriation du contexte : enjeux et acteurs	26
3.3.	Problématisation et hypothèse	26
3.4.	L’enquête par entretien	27
4.	Tâche 03 – Implémentation de Casses.....	27
4.1.	Données utilisées	29
4.2.	Modèles utilisés.....	29
4.3.	Résultats	30

4.3.1.	LEYP	30
4.3.2.	LEYP2s.....	34
5.	Tâche 04 – Intégration des outils dans le système d’information	36
5.1.	Choix stratégiques	36
5.2.	Données historiques.....	36
5.2.1.	Données sur les défaillances originellement non affectées aux ID de tronçons.....	36
5.2.2.	Données complétées sur les tronçons Hors Service	36
5.2.3.	Renouvellements de canalisations	36
5.3.	Données vivantes	37
5.3.1.	Covariables additionnelles des tronçons.....	37
5.3.2.	Fichier des casses	38
6.	Tâche 05 – Exploitation des données de la sectorisation	38
6.1.	Caractéristiques des cellules et covariables des tronçons	38
6.1.1.	Mise à jour du calcul des pressions	38
6.1.2.	Corrosivité de l’eau.....	38
6.1.3.	Autres covariables	39
6.2.	Stabilité des cellules dans le temps.....	39
6.2.1.	Problématique	39
6.2.2.	Tronçons et cellules.....	39
6.2.3.	Démarche adoptée pour la validation de l’historique d’une cellule.....	39
6.2.4.	Regroupements de cellules	40
6.3.	Calcul du débit des cellules	40
6.4.	Caractérisation des cellules de distribution	41
6.4.1.	Linéaire des cellules.....	41
6.4.2.	Nombre de branchements	41
6.4.3.	Densité de branchements	42
6.4.4.	Nombre de compteurs par cellule ou groupement de cellule	43
6.4.5.	Volume journalier.....	44
6.4.6.	Typologie des cellules.....	44
7.	Tâche 06 – Analyse des pratiques de recherche de fuites	44
8.	Tâche 07 – Modélisation de la probabilité d’un tronçon d’être fuyard	46
8.1.	Principe de la modélisation	46
8.2.	Données d’entrée de la régression logistique (GLM) de l’état fuyard	47
8.3.	Résultats	48

8.3.1.	Succès des campagnes RF	48
8.3.2.	Régression logistique.....	50
8.4.	Objectifs futurs.....	57
8.4.1.	Amélioration de l’hypothèse d’inspection des tronçons	57
8.4.2.	Exploration de la prédiction de la probabilité d’état fuyard.....	57
8.4.3.	Probabilité d’état fuyard non conditionnée par l’inspection.....	57
9.	Tâche 08 – Analyse de la vulnérabilité des usagers	58
9.1.	Contexte et problématique	58
9.2.	Construction du questionnaire.....	59
9.2.1.	Le scénario « Restriction des usages »	60
9.2.2.	Le scénario « Performance du service rendu »	61
9.2.3.	Structure du questionnaire	61
9.3.	Protocole d’enquête et collecte des données.....	62
10.	Tâche 13 – Coordination du Projet et valorisation.....	64
10.1.	Animation	64
10.2.	Moyens humains affectés par INRAE au projet.....	64
10.3.	Valorisations	64
10.4.	Planning prévisionnel ajusté.....	65
11.	Conclusion	65
12.	Références.....	67
13.	Annexes	69

Figures

Figure 1 : Illustration du regroupement en campagnes pour une cellule.....	24
Figure 2 : Représentation des périodes de calage et de validation lors du calage du modèle.....	28
Figure 3 : Représentation des périodes de calage et de prédiction pour un objectif de prédiction des défaillances.....	28
Figure 4 : Courbe de validation des fontes du modèle LEYP.....	32
Figure 5 : Courbe de validation de l'acier du modèle LEYP.....	32
Figure 6 : Courbe de validation de l'AMCI + autres matériaux du modèle LEYP.....	32
Figure 7 : Extrait de prédiction des défaillances à Tournai.....	33
Figure 8 : Courbe de validation des fontes du modèle LEYP2s.....	35
Figure 9 : Courbe de validation de l'acier du modèle LEYP2s.....	35
Figure 10 : Courbe de validation de l'AMCI + autres matériaux du modèle LEYP2s.....	35
Figure 11 : Renouvellements et extensions du réseau.....	37
Figure 12 : Caractérisation du facteur agressif de l'eau distribué par la SWDE (fichier SWDE : Equilibre calco 2022 Complet.xlsx).....	38
Figure 13 : Répartition des cellules en fonction de la longueur de réseau.....	41
Figure 14 : Répartition des cellules en fonction du nombre de branchements.....	42
Figure 15 : Nombre de logements par branchement.....	42
Figure 16 : Répartition des cellules par densité de branchement.....	43
Figure 17 : Nombre de compteurs par cellule.....	43
Figure 18 : Répartition des cellules en fonction du nombre de débits journaliers calculés.....	44
Figure 19 : Nombre de campagnes de RF par secteur de distribution entre 2016 et 2021.....	49
Figure 20 : Nombre de campagnes de RF par année sur toute la SWDE.....	49
Figure 21 : Odd-ratios des coefficients du modèle fixe et leurs intervalle de confiance à 95%.....	53
Figure 22 : Odd ratios des coefficients du modèle mixte et leurs intervalle de confiance à 95%.....	54
Figure 23 : Distribution de la probabilité d'état fuyard des tronçons au 1 ^{er} janvier 2022 issue du modèle mixte.....	55
Figure 24 : Probabilité d'état fuyard selon le pourcentage d'effectif cumulé de tronçons.....	56
Figure 25 : Probabilité d'état fuyard selon le pourcentage d'effectif cumulé de tronçons. Zoom.....	56
Figure 26 : Probabilité d'état fuyard selon le pourcentage de longueur cumulée des tronçons.....	56
Figure 27 : Probabilité d'état fuyard selon le pourcentage de longueur cumulée des tronçons. Zoom.....	56
Figure 28 : exemple de carte de choix du scénario « Restrictions des usages ».....	60
Figure 29 : exemple de carte de choix du scénario « Performance du service fourni ».....	61
Figure 30 : mail envoyé par la SWDE à un échantillon aléatoire de clients habitant dans les communes d'enquête.....	63
Figure 31 Planning prévisionnel ajusté au 29/07/2022.....	65
Figure 32 : Distribution des distances (mètres) entre conduite portant la défaillance et la conduite la plus proche de la cellule inscrite dans les fichiers fuites (graphe du bas : zoom).....	72
Figure 33: Illustration de la complétion de DDP en fonction du matériau par la méthode c2 sur un réseau fictif.....	91
Figure 34 : Performance de la complétion de DDP sur le secteur de Mons (oubli par ICAR de rue)....	93
Figure 35 : Impact de l'oubli et de la complétion de 40 % des matériaux sur l'aire sous la courbe de validation du LEYP.....	96

- Figure 36 : Impact de l'oubli et de la complétion de 40 % des matériaux sur le coefficient α du LEYP.97
- Figure 37 : Impact de l'oubli et de la complétion de 40 % des matériaux sur les paramètres du LEYP.97
- Figure 38 : Impact de l'oubli et de la complétion de 40 % des matériaux sur les rangs des tronçons classés par risque de défaillance par le LEYP. 97
- Figure 39 : Indicateur de délai-seuil inter-prestations optimal pour les données des 17 secteurs. ... 104

Tableaux

Tableau 1 : Tâches du projet GePaME.....	12
Tableau 2 : Répartition des anomalies sur les différents éléments du réseau.	16
Tableau 3 : Dates utilisées pour la localisation et la réparation des défaillances.....	18
Tableau 4 : Indicateur d'incertitude pour chaque défaillance de conduite affectée aux tronçons du SIG.....	19
Tableau 5 : Bilan des défaillances de conduites affectées aux tronçons du SIG.....	20
Tableau 6 : Détail du nombre d'affectations par méthode et par année.	21
Tableau 7 : Estimation des paramètres du modèle LEYP pour les strates Fontes, Acier et AMCI + autres matériaux.	30
Tableau 8 : Estimation des paramètres du modèle LEYP2s pour les strates Fontes, Acier et AMCI + autres matériaux.	34
Tableau 9 : Indicateur de corrosivité.....	38
Tableau 10 : Nombre de cellules suivant leur période de stabilité.....	40
Tableau 11 : Groupements de cellule et leur période de stabilité.....	40
Tableau 12 : Liste des acteurs interviewés (janvier-mai 2022)	45
Tableau 13 : Liste des observations réalisées (mars-mai 2022).....	46
Tableau 14 : Nombre de campagnes de RF selon leurs résultats.	49
Tableau 15 : Résultat des campagnes de RF.	50
Tableau 16 : Nombre de campagnes de RF selon leurs résultats.	50
Tableau 17 : Comparaison des estimations des paramètres entre modèle fixe et mixte de régression logistique sur la probabilité d'état fuyard.....	51
Tableau 18 : Liste des communes d'enquête.....	59
Tableau 19 Personnels non-permanents et stagiaires INRAE	64
Tableau 20 : Taux de désaccord entre cellules selon l'année.	69
Tableau 21 : Taux de désaccord entre cellules par secteur.	69
Tableau 22 : Désaccords de cellules fréquents pour certaines cellules.....	70
Tableau 23 : Distribution des distances (mètres) entre conduite portant la défaillance et la conduite la plus proche de la cellule inscrite dans les fichiers fuites.....	71
Tableau 24 : Colonnes du fichier des fuites concaténé.....	74
Tableau 25 : Expressions régulières utilisées pour extraire les matériaux.	76
Tableau 26 : Taux de succès d'affectation par méthode.....	81
Tableau 27 : Conduites ES corrigées.	82
Tableau 28 : Conduites HS corrigées.....	86
Tableau 29 : Calage du modèle LEYP de référence sur Couillet.....	94
Tableau 30 : Délai-seuil optimal selon le secteur.....	104

Liste des sigles et abréviations

Abréviation	Objet	Description
AEP	Alimentation en Eau Potable	
AMCI	Amiante Ciment	
ASTEE	Association Scientifique et Technique pour l'Eau et l'Environnement	Association française de professionnels de l'eau et des déchets.
BAT	Béton Âme Tôle	Matériau de conduite.
BDD (BD)	Base De Données	
BI/PI	Borne Incendie / Poteau Incendie	Borne, Bouche, ou Poteau Incendie (BI/PI). Synonyme d'hydrant.
CILE	Compagnie Intercommunale Liégeoise des Eaux	
CPS	Comité de Pilotage Stratégique	Réunion annuelle d'orientation et de suivi du projet.
CR	Compte Rendu	
CSO	Comité de Suivi Opérationnel	Réunion périodique de suivi du projet.
CLC	Corine Land Cover	Occupation du sol.
CT	Court Terme	
DDC	Date De Casse	Date à laquelle l'élément du réseau AEP (comme la conduite) a subi une défaillance.
DDP	Date De Pose	Date à laquelle la conduite a été posée.
DE	Diamètre Extérieur	Diamètre extérieur (mm) d'un tuyau ou d'une conduite.
DHS	Date de mise Hors Service	Date à laquelle la conduite a été mise hors service, ou abandonnée. La DHS de la canalisation est une donnée nécessaire à l'implémentation du modèle LEYP2s.
DIA	Diamètre	Selon le contexte, peut désigner DN ou DE.
DN	Diamètre Nominal	Diamètre nominal (mm) d'un tuyau ou d'une conduite.
EPSG	European Petroleum Survey Group	
ES	En Service	Utilisé pour décrire des conduites qui sont encore utilisées à l'heure actuelle pour l'AEP.
FD	Fonte Ductile	
FG	Fonte Grise	
FME	Feature Manipulation Engine	Logiciel de type ETL (Extraction, Transformation, Chargement) utilisé par la SWDE pour automatiser certaines opérations de SIG.
GLM	Generalized Linear Model	Modèle Linéaire Généralisé, extension de la théorie des modèles linéaires.
GLMM	Generalized Linear Mixed Model	Extension des GLM pour prendre en compte la non-indépendance des données (prise en compte d'effets aléatoires en sus d'effets fixes).
GPI	Gestion Patrimoniale des Infrastructures	

HMT	Hauteur Manométrique Totale	Grandeur en mètres de colonne d'eau calculée pour approximer la charge hydraulique en régime stationnaire dans les conduites.
HS	Hors Service	Synonyme de « abandonné », utilisé pour les conduites hors service.
ICAR	Inventaire Centralisé des Adresses et des Rues en Wallonie	Système utilisé par le SPW pour identifier de façon unique chaque rue ou point d'adresse par un code alphanumérique.
ID	Identifiant	Code utilisé dans un fichier pour identifier de façon unique chaque objet. L'OT SAP est un type d'ID.
IDT	Identifiant de Tronçon	
INRAE	Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement	
INS (code)	Institut National de la Statistique	Code à 5 caractères propre à chaque commune en Belgique.
IWA	International Water Association	Réseau de professionnels du monde de l'eau.
LESAM	Leading-Edge Strategic Asset Management	Conférence internationale organisée par des membres de l'IWA sur le thème de la GPI.
LEYP	Linearly Extended Yule Process	Modèle probabiliste à base du processus de Yule utilisé pour la prédiction des défaillances (fuites et casses) des conduites.
LT	Long Terme	
MAG	Mons Ath Gaurain	Regroupement de 3 secteurs de distribution utilisé pour étendre les méthodes développées sur Mons.
MAT	Matériau	Matériau de l'élément du réseau AEP.
MNT	Modèle Numérique de Terrain	Représentation numérisée de la topographie du terrain de surface.
NA	Not Available	Code indiquant une donnée non disponible ou non applicable.
NPS	Nominal Pipe Size	Diamètre de tuyau ou de conduite exprimé en pouces.
OT	Ordre de Travail	Numéro utilisé dans SAP pour identifier de façon unique chaque intervention.
PE	Polyéthylène	
PI	Poteau Incendie	
PICC	Projet Informatique de Cartographie Continue	Cartographie des bâtiments de la Wallonie.
PoWalCo	Portail Wallon de Coordination	Portail internet permettant l'échange d'informations sur les intentions de travaux, la coordination des chantiers et la gestion des autorisations en Wallonie.
PVC	Polychlorure de Vinyle	
QGIS	Quantum Geographic Information System	Logiciel libre de SIG.
RF	Recherche de Fuite	

SAP	S ystems A pplications and P roducts in D ata P rocessing	Progiciel de gestion intégré utilisé par la SWDE notamment pour archiver les interventions.
SCR	S ystème de C oordonnées de R éférence	Informations mathématiques sur la méthode de projection de géométrie sphérique sur un plan.
SI	S ystème d' I nformation	Système (logiciel) utilisé pour stocker et traiter l'information.
SIG	S ystème d' I nformation G éographique	Système (logiciel) utilisé pour stocker et traiter l'information géographique.
SPW	S ervice P ublic de W allonie	Administration publique Wallonne.
SRRE	S chéma R égional des R essources en E au	Outil de planification et de réglementation de l'exploitation des ressources en eau impliquant la coordination des actions et des investissements des différents opérateurs du domaine de l'eau en Wallonie.
SWDE	S ociété W allonne D es E aux	
TP	T rop P lein	Dispositif d'évacuation de l'eau arrivant en excès dans un réservoir.
TSM	T echniques S ciences M éthodes	Journal sur les thèmes de l'eau et de l'environnement géré par l'ASTEE.
ZDE	Z one de D istribution d' E au	Voir glossaire.
ZF	Z one F onctionnelle	Voir glossaire.

Glossaire

Terme employé	Signification
Casses (C majuscule)	Logiciel de prédiction des défaillances de conduites.
casse (c minuscule)	Synonyme de crash.
Cellule	Ensemble continu de conduites délimité par des vannes fermées et/ou des compteurs de sectorisation (« secteur hydraulique » en France).
Centre d'exploitation (CED)	Secteur de type distribution.
Crash	Défaillance pour laquelle il existe une indication forte, souvent visuelle comme un écoulement en surface, mais parfois indirecte comme la chute rapide d'un niveau de réservoir. En terminologie courante, correspond à une « fuite signalée ».
Défaillance	Fuite ou crash (casse) d'un élément du réseau (conduite, raccordement, appareil, hydrant...).
Défaillance de conduite	Fuite ou crash au type d'activité SAP concernant les interventions sur conduites (105, 999, voir 2.1.4.2).
Fichier des casses	Un des fichiers d'entrée du logiciel Casses. Concerne les défaillances de conduites.
Fuite	Défaillance non crash. En terminologie courante, correspond à une « fuite détectée par une opération de recherche active ».
Hydrant	Borne, Bouche, ou Poteau Incendie (BI/PI).
Secteur	Ensemble de cellules gérées en commun. Il existe quatre secteurs de production (A, B, C, D) et 17 secteurs de distribution.
Raccordement	Ensemble d'éléments de liaison entre la conduite principale et le compteur d'un abonné (« branchement » en France).
Tronçon	Synonyme de conduite. Ensemble continu de tuyaux aux caractéristiques identiques. Élément de description du réseau au sens du SIG de l'exploitant.
Zone de Distribution d'Eau (ZDE)	Zone géographique dans laquelle les eaux destinées à la consommation humaine proviennent d'une ou de plusieurs sources et à l'intérieur de laquelle la qualité est considérée comme uniforme (synonyme parfois trouvé : aire de distribution, « unité de distribution » en France).
Zone Fonctionnelle (ZF)	Groupe de conduites identifié pour mieux caractériser le chemin de l'eau depuis son captage jusqu'au robinet de l'utilisateur. Utilisé pour l'analyse des risques sur le réseau notamment grâce à l'absence du découpage production/distribution et grâce à une taille plus pertinente que les ZDE (permettant de prendre en compte leurs interconnexions).

1. Introduction

1.1. Contexte

SWDE et INRAE ont décidé de s'associer pour conduire des travaux de recherche dans le domaine la gestion patrimoniale des réseaux d'eau potable. Ils ont ainsi conclu, le 9 juin 2020, un contrat de recherche qui établit une collaboration pour mener à bien le projet « Gestion Patrimoniale Multi-Echelles des réseaux d'eau potable » (GePaME). Le projet GePaME, d'une durée prévisionnelle de trois ans, a démarré le 1^{er} septembre 2020. Le contrat prévoit que les résultats du projet soient présentés dans des rapports annuels établis par l'INRAE. Un premier rapport intermédiaire a été produit en septembre 2021 (Rodriguez et al., 2021).

Le présent rapport est le second rapport intermédiaire du projet GePaME, il retrace les travaux réalisés et les résultats obtenus au cours de la deuxième année de réalisation du projet, c'est-à-dire entre septembre 2021 et août 2022.

1.2. Contenu du projet GePaME

Le projet GePaME est organisé en trois volets :

- Volet 1 : compréhension du processus de fuite et efficacité de la recherche de fuite (RF) ;
- Volet 2 : aide à la sélection multi-objectifs et contrainte des chantiers de renouvellement ;
- Volet 3 : simulation long terme de stratégies de gestion patrimoniale.

Il se décompose en treize tâches opérationnelles (Tableau 1) qui doivent aboutir à des avancées scientifiques valorisées par l'INRAE et à des outils d'aide à la décision qui ont vocation à être intégrés dans les pratiques et dans le système d'information de la SWDE.

Tableau 1 : Tâches du projet GePaME.

Tâche	Dénomination	Référent SWDE	Référent INRAE
01	Prise en main et traitement des données	F. Haan	A. Husson
02	Analyse de la gouvernance de la SWDE en matière de GPI	D. Brunet	K. Caillaud
03	Implémentation de Casses	J-G. Baudoin	Y. Le Gat
04	Intégration des outils dans le système d'information	J-G. Baudoin	Y. Le Gat
05	Exploitation des données de la sectorisation	D. Devos	A. Husson
06	Analyse des pratiques de recherche de fuites	O. Pinsar	K. Caillaud
07	Modélisation de la probabilité d'un tronçon d'être fuyard	O. Pinsar	Y. Le Gat
08	Analyse de la vulnérabilité des usagers	D. Brunet	B. Rulleau
09	Modélisation statistique couplée des fuites et des crashes	O. Pinsar	Y. Le Gat
10	Outil multicritère de programmation court terme	D. Brunet	Y. Le Gat
11	Efficacité des actions de recherche de fuite	O. Pinsar	E. Renaud
12	Etude de scénarios long terme et simulation	D. Brunet	Y. Le Gat
13	Coordination du Projet et valorisation	D. Brunet	E. Renaud

1.3. Rappel des avancées de la première année du projet

Les avancées de la première année du projet (septembre 2020 – août 2021) ont principalement concerné la préparation des données et le cadrage méthodologique des tâches amont. Ainsi les résultats obtenus ont essentiellement concerné les tâches 01 « Prise en main et traitement des

données » et 05 « Exploitation des données de la sectorisation » et des choix méthodologiques ont été proposés pour les tâches 02 « Analyse de la gouvernance de la SWDE en matière de GPI » et 07 « Modélisation de la probabilité d'un tronçon d'être fuyard » (Rodriguez et al., 2021).

A l'issue de ce travail, les sources pertinentes de données ont été identifiées et expertisées (SIG, SAP, Fichiers « Fuite », données de télégestion « Perf'O », ressources géographiques publiques de Wallonie). Pour combler l'écart entre la forme sous laquelle l'information est disponible et les besoins de l'étude, en particulier la nécessité de relier des événements et des caractéristiques aux tronçons et aux cellules, des méthodes et des codes de calcul ont été développés à partir des données du secteur de Mons. Les premières versions des méthodes opérationnelles permettant de produire les fichiers d'entrée du logiciel Casses et de calculer les caractéristiques nécessaires à l'établissement d'une typologie des cellules ont été produites. Parallèlement, des approches méthodologiques ont été construites pour servir de fil conducteur à la réalisation des tâches 02 « Analyse de la gouvernance de la SWDE en matière de GPI » et 07 « Modélisation de la probabilité d'un tronçon d'être fuyard ».

1.4. Contenu du rapport

Le présent rapport intermédiaire 2022 concerne les tâches 1 à 8 et 13 (Tableau 1), un chapitre étant consacré à chacune de ces tâches. Les principaux aspects abordés sont les suivants :

- Tâche 1 « Prise en main et traitement des données » : passage à l'échelle de l'ensemble du territoire de la SWDE, améliorations de la méthode d'affectation des défaillances aux tronçons, fiabilité de l'affectation, complétion des données manquantes des caractéristiques principales des tronçons, traitement des données de recherche de fuites ;
- Tâche 2 « Analyse de la gouvernance de la SWDE en matière de GPI » : contexte, méthodologie mise en œuvre, premiers retours sur les entretiens ;
- Tâche 3 « Implémentation de Casses » : modèles et résultats des prévisions de défaillances sur les secteurs de Mons, Ath et Gaurain, prise en compte de la fiabilité des données, cartographie des résultats ;
- Tâche 4 « Intégration des outils dans le système d'information » : arbitrages effectués par le sous-groupe de travail dédié, premiers transferts des méthodes de calcul de caractéristiques des tronçons ;
- Tâche 5 « Exploitation des données de la sectorisation » : collecte et traitement des données de débit, prise en compte des variations temporelles des périmètres des cellules, compléments et améliorations des méthodes de calcul des indicateurs associées aux cellules ;
- Tâche 6 « Analyse des pratiques de recherche de fuites » : méthodologie mise en œuvre et premiers retours sur l'enquête en immersion ;
- Tâche 7 « Modélisation de la probabilité d'un tronçon d'être fuyard » : principes et objectifs des modèles, résultats des modélisations réalisées sur Mons, Ath et Gaurain ;
- Tâche 8 « Analyse de la vulnérabilité des usagers » : méthodologie mise en œuvre, contenu du questionnaire adressé aux usagers, point sur le déroulement de l'enquête.
- Tâche 13 « Coordination du Projet et valorisation » : animation, moyens humains, valorisations, planning.

2. Tâche 01 – Prise en main et traitement des données

NB : dans le reste du rapport, sauf précisé autrement, « secteur » désignera « secteur de distribution » (voir glossaire en début de document), et « défaillance » pourra concerner différents éléments du réseau associés à des types d'activités SAP différents. « fuite » possède une signification précise qui n'est pas nécessairement appliquée en dehors du contexte du projet. Nous incitons les lecteurs à lire attentivement le glossaire pour éviter également la confusion entre plusieurs termes similaires aux significations différentes dans le contexte de ce projet (p. ex., casse vs Casses vs fichier des casses).

2.1. Génération des fichiers pour Casses sur l'ensemble du territoire SWDE

2.1.1. Rappel du contexte

Le rapport intermédiaire de 2021 détaille la méthode employée pour produire les fichiers pour Casses (fichier des tronçons et fichier des casses) sur le secteur de Mons (5230), qui représente environ 2 % du linéaire total du réseau SWDE et 3 % du total des interventions (2 % des réparations de conduites). Pour rappel, une partie importante du travail a consisté à affecter les défaillances aux tronçons « agrégés » (regroupés) du SIG, grâce à un degré de connaissance variable de la localisation de chaque défaillance (point SIG, ou rue avec/sans numéro d'immeuble) et éventuellement de ses caractéristiques (MAT et/ou DIA de la conduite réparée). Ces éléments de connaissance sont renseignés sous forme de tables dans plusieurs fichiers de la SWDE généralement liés entre eux par un Ordre de Travail (OT). Les défaillances affectées aux ID de tronçons et de type d'activité SAP égal à 105 (« réparation de conduite ») ont alors servi de base pour le fichier des casses.

Depuis, l'ensemble des données nécessaires pour faire ce travail sur la totalité des 17 secteurs de distribution (= centres d'exploitation, voir glossaire) a été transmis :

- Liste des cellules et leur secteur d'appartenance (ainsi que ZDE et ZF).
- Fichiers « fuites » (Excel).
- Extraits SAP.
- Réseau SIG portant les défaillances (conduites ES et HS, hydrants, et vannes).
- « Anomalies SIG » et « opérations SIG » (voir 2.1.4.2).
- Pièces de réparation associées aux OT de SAP.

Pour rappel, le projet GePaME utilisera les données collectées par la SWDE jusqu'au 31/12/2021. Pour les traitements appliqués à ces données sur les 17 secteurs, l'année 2021 est achevée (et archivée) et ses données peuvent être prises en compte, ce qui n'était pas encore le cas pour le secteur de Mons. Quelques adaptations ont été apportées à la méthode d'affectation des défaillances aux conduites en analysant les résultats de ses étapes intermédiaires sur d'autres secteurs que Mons.

2.1.2. Prise en compte des types d'activités pertinents pour les pertes en eau

L'affectation des défaillances aux tronçons du SIG concerne en premier lieu le numéro d'activité SAP 105 (réparation de conduite) pour la génération des fichiers pour Casses.

Mais les pertes en eau concernent aussi d'autres types d'activité SAP, comme le numéro 92 (réparation de raccordement), et encore les numéros 10, 11, 13, 30, 31, 32, 50, 72, 82, 100, 112, 113, 114, 180, et 185, pour les OT qui sont présents dans les fichiers fuites (voir rapport intermédiaire 2021). Il est important de noter que pour les numéros d'activité autres que 105, les informations de matériau et/ou de diamètre provenant des fichiers fuites et/ou des pièces de réparation ne sont pas disponibles ou pas utilisables (p.ex. lorsque le matériau du raccordement est indiqué dans les fichiers fuites). Ces numéros d'activités sont cependant pertinents pour les autres tâches du projet (notamment les tâches

05 et 07). L'archive des défaillances historiques construite grâce à l'affectation des défaillances aux tronçons est similaire en format au fichier des casses, mais inclut tous les types d'activités pertinents ci-dessus.

2.1.3. Adaptations suite à l'application sur les secteurs de Mons, Ath, et Gaurain

Suite au développement des méthodes sur le secteur de Mons, il a été décidé conjointement avec la SWDE (CSO du 07/06/2021) d'étendre les traitements sur une zone de test plus large, composée de 3 secteurs de Mons (5230), Ath (5110), et Gaurain (5120), abrégé MAG, représentant 17 % du linéaire du réseau de distribution et 17 % des interventions (14 % des réparations de conduite). Les paragraphes suivants détaillent les adaptations de la méthode d'affectation des défaillances aux conduites justifiées par des analyses sur MAG.

2.1.3.1. Élargissement des plages de DN déduites des pièces de réparation

En analysant au cas par cas les résultats de l'affectation des défaillances de conduites (type d'activité 105) aux tronçons du SIG sur MAG, nous avons pu constater que certaines plages de DN déduites des pièces de réparation et utilisées pour filtrer les conduites candidates étaient trop courtes ou incompatibles avec les conduites présentes à la localisation renseignée. Nous avons donc décidé tout d'abord d'utiliser un nombre plus restreint de pièces de réparation (en omettant les joints plats, les vannes et accessoires, etc.) qui pourraient fausser la sélection de la conduite (par exemple dans le cas d'une pièce correspondant à une vanne, qui aura souvent un DN inférieur au DN de la conduite sur laquelle elle est posée). Les seules pièces finalement utilisées pour déduire le DN de la conduite réparée sont donc celles posées à même la conduite : les prises en charge, les manchons (bague, coulissant, de liaison, de réparation), les tuyaux et allonges, les brides d'emboîtement, les coudes, tés, réductions, les raccords, et les plaques pleines. De plus, la méthode d'extraction des DN à partir des pièces a été modifiée comme suit :

- Pour les pièces avec plage de diamètre externe (DE), au lieu de prendre pour DN min de la conduite le premier DN du SIG immédiatement inférieur au DE min de la pièce, c'est le deuxième DN du SIG immédiatement inférieur au DE min qui est choisi. De plus, au lieu de prendre pour DN max de la conduite le premier DN du SIG immédiatement inférieur au DE max, c'est le deuxième DN du SIG immédiatement supérieur au DE max qui est choisi.
- Pour les pièces qui donnent explicitement le DN associé de la conduite (pour lesquelles la plage de diamètre externe était ignorée), on ne se limite plus à ce dernier, mais on prend comme plage DNmin-DNmax les valeurs entre le premier DN du SIG immédiatement inférieur au DN indiqué par la pièce et le premier DN du SIG immédiatement supérieur au DN indiqué par la pièce.

Ces deux modifications permettent de considérablement élargir la plage de DN compatibles lors de la sélection de conduites. Par exemple, avec l'ancienne méthode, la pièce "PRISE EN CHARGE DN125 130-160MM FI-DN1" était associée à (DNmin, DNmax) = (125, 125), qui est trop étroite ; désormais on a (DNmin, DNmax) = (110, 150), ce qui prend en compte les incertitudes éventuelles de la pièce ou des diamètres du SIG. Par ailleurs, avec l'ancienne méthode, la pièce "PIECE REPA; INOX(D±5°) 87-109MM" était associée à (DNmin, DNmax) = (80, 100) ; désormais on a (DNmin, DNmax) = (75, 125).

2.1.3.2. Ajout du critère de la cellule d'appartenance dans la sélection des conduites

Une analyse a posteriori des affectations des défaillances (tous types d'activités pertinents) aux conduites de MAG a permis de constater que 17 % des défaillances affectées dont la cellule est connue grâce aux fichiers fuites (880 cas sur 5256) sont localisées sur des conduites du SIG qui portent une autre cellule que celle mentionnée dans les fichiers fuites. L'analyse détaillée (présentée en Annexe 1) a révélé que le numéro de cellule indiqué dans les fichiers fuites est partiellement utile pour trouver la bonne conduite pour chaque défaillance, mais qu'il peut aussi parfois être imprécis. Ainsi, nous avons décidé d'inclure le critère d'appartenance à la cellule mentionnée dans les fichiers fuites pour les conduites candidates à l'affectation des défaillances, mais de façon non obligatoire (i.e., le critère

est ignoré s'il élimine toutes les conduites candidates). De plus, en termes de priorité de sélection de conduite, ce critère **intervient après les filtres sur les caractéristiques de la conduite** (DIA et/ou MAT, propres aux défaillances de conduite seulement), **et avant le choix par défaut** de la conduite la plus proche de la défaillance (adresse renseignée avec le numéro d'immeuble) ou de la conduite la plus longue compatible (adresse sans numéro d'immeuble).

2.1.4. Autres adaptations de la méthode

2.1.4.1. Ajustement des traitements des fichiers fuites

La méthode de concaténation des fichiers fuites a été ajustée suite à des petites différences de format de données entre secteurs. L'Annexe 2 détaille la méthode finale, utilisée de façon identique sur chacun des 17 secteurs. De façon simplifiée, par rapport à Mons, le fichier des fuites concaténé pour les 17 secteurs contient quelques nouvelles colonnes (agent n°2, permalog fixe vs mobile, utilisation de gaz traceur, utilisation de données satellitaires), et la méthode finale formate de façon plus robuste les numéros de cellule, les diamètres, les matériaux, les OT (réparation et systématique) et les dates de localisation et de réparation des défaillances. Enfin, une correction automatique des doublons d'OT présents dans les fichiers fuites a été développée et appliquée.

2.1.4.2. Gestion des « anomalies SIG » et « opérations SIG »

Pour rappel, les « anomalies SIG » utilisées dans le projet sont une couche Shapefile de points (X, Y) localisant précisément sur les conduites du SIG (ou à quelques centimètres près) les défaillances observées par les équipes de terrain de la SWDE. Les anomalies SIG concernent des interventions sur divers équipements du réseau (Tableau 2).

Tableau 2 : Répartition des anomalies sur les différents éléments du réseau.

Type d'activité SAP de l'anomalie SIG	Équipements concernés	Nombre d'anomalies SIG (% du total)
100 ou 105	conduite	8 493 (54 %)
82 ou 92	raccordement	3 385 (21 %)
11 ou 13	hydrant	1 143 (7 %)
113 ou 114	appareil	665 (4 %)
autres	divers	174 (1 %)
inconnu	inconnu	2 004 (13 %)
TOTAL		15 864

Pendant la rédaction du rapport intermédiaire 2021 (juillet 2021), la SWDE a signalé un autre type de point SIG pouvant être utile pour le projet, appelé « opération SIG », et représentant les réparations de conduites pour lesquelles il n'y a pas d'anomalie et remontées au SIG grâce à un croquis de pose (représentant donc essentiellement les crashes). Ces opérations sont similaires en format aux anomalies, et indiquent donc immédiatement quelle conduite (désignée par son ID) est concernée par la défaillance. Une inspection des fichiers des anomalies et opérations a révélé le besoin de réaliser deux traitements supplémentaires pour leur prise en compte dans les fichiers de casses :

- Supprimer les anomalies qui sont des doublons d'opérations (en conservant en priorité les opérations, qui sont plus précises en géolocalisation et en date).

- Prendre en compte les anomalies et opérations sans OT renseigné, sans créer de doublons avec les défaillances de SAP éventuellement correspondantes.

Concernant le premier point, nous avons simplement supprimé les anomalies possédant le même OT qu'une opération (200 anomalies éliminées). Les anomalies et opérations sans OT ont été isolées de l'affectation des défaillances aux conduites (l'ID du tronçon est déjà connu), et ajoutées en fin de traitements comme des défaillances avec un OT défini à la valeur 0. Pour éviter de créer des éventuels doublons avec des défaillances correspondantes et possédant un OT renseigné dans SAP, une vérification a été effectuée a posteriori :

- Les défaillances avec OT non nul qui ont été attribuées par la rue avec ou sans numéro d'immeuble ont vu leur ID de conduite et leur date de casse corrigés par ceux fournis dans un point SIG avec OT nul si ce dernier concernait la même rue et une date de réparation distante de 4 jours au maximum (laissant une marge d'erreur sur les dates) de celle de l'OT non nul (1 059 corrections d'ID apportées). En cas de conflit entre une anomalie et une opération SIG pour la correction, la priorité est donnée à l'opération.
- Les défaillances avec OT nul (= vide à l'origine) utilisées pour corriger les ID de conduite ont alors été supprimées, et celles qui restaient (881 cas) ont été conservées comme des défaillances valides. **Un nouveau type d'activité leur a été affecté : 999 (« point SIG sans OT »), qui est pris en compte de la même manière que le type d'activité 105 dans la génération du fichier des casses.** Ceci part de l'hypothèse qu'il s'agit en majorité de réparations de conduites, ce qui est systématiquement vrai pour les opérations. En effet, il apparaît qu'environ 60 % des anomalies avec OT concernent des réparations de conduites (contre seulement 20 % pour les réparations de raccordements par exemple).

Enfin, nous avons aussi constaté que 2562 anomalies SIG n'avaient pas leur ID de conduite renseigné (ce qui n'était pas le cas sur les seuls secteurs de MAG). De plus, nous avons observé des mélanges entre l'objet SIG portant l'anomalie (conduite ou hydrant ou vanne) et le type d'activité de l'anomalie (réparation de conduite ou d'hydrant ou de vanne). Par exemple on trouve 500 anomalies portées par des conduites mais concernant des BI/PI. Afin de corriger ces deux problèmes simultanément, nous avons récupéré de façon systématique l'ID de conduite la plus proche de chaque point d'anomalie, dans la limite de 1 000 m (évitant quelques incohérences). De plus, lorsque l'ID de conduite était déjà renseigné pour une anomalie concernant une conduite en service ou abandonnée, ce dernier a été conservé et non remplacé (moins de 300 cas).

2.1.4.3. Gestion des défaillances sans panne, archivées, ou doublons

Les défaillances archivées dans SAP sont associées à des codes et textes de cause (PM-XXX) et de panne (YYY), comme PM-D100-100 pour une « Fuite sur Réducteur de pression ». Nous avons remarqué environ 56 000 OT associés à des codes indiquant qu'il n'y a pas de panne. Les OT correspondants ont été éliminés des traitements. Il s'agit des codes suivants :

- PM-C310-100 : « Pas de panne ».
- PM-D999-999 : « Pas de panne ».
- PM-FUITE-999 : « Pas de fuite ».

Un archivage des OT anciens (principalement ceux datant de 2011) a eu lieu peu de temps avant l'envoi par la SWDE des extraits SAP sur tous les secteurs. Cet archivage a rendu impossible l'extraction d'un certain nombre d'OT d'avant 2011, ce qui n'a pas d'impact sur le projet étant donné qu'il a été décidé de définir la période d'observation des défaillances à partir de 2011 qui marque le début de l'enregistrement exhaustif des défaillances dans le SI de la SWDE. Les OT de 2011 archivés ont tous pu être retrouvés par la SWDE, néanmoins avec quelques informations en moins par rapport aux OT non archivés. Pour traiter les OT archivés de la même manière que les autres, il a été nécessaire de redéfinir le type d'OT à partir du premier chiffre de l'OT (CLT pour 5, DIST pour 6 sachant que le type RFU

n'existait pas en 2011). De plus, les OT archivés n'ont pas de « date de début planifié » ni de « date de début réel » mais une « date de fin de panne » qui a été utilisée en substitution.

Enfin, une suppression des défaillances doublons a été effectuée. Premièrement, près de 14 000 doublons ont été identifiés et éliminés dans les extraits bruts de défaillances dans SAP. Deuxièmement, nous avons observé environ 1 600 OT qui, après affectation de la défaillance à une conduite du SIG, donnaient lieu à la même information {ID de conduite, date de casse, type d'activité SAP} qu'un autre OT. Il s'agit donc de mêmes défaillances encodées de plusieurs manières dans le SI de la SWDE. La suppression de ces doublons a été effectuée en conservant la copie de la défaillance dont les informations de géolocalisation étaient les plus précises, c'est-à-dire, dans l'ordre de priorité suivant :

1. Défaillance déjà connue dans le SIG (anomalie ou opération).
2. Défaillance attribuée par rue avec numéro d'immeuble.
3. Défaillance attribuée par la rue seule.

De plus, en cas d'égalité de priorité au point précédent, et uniquement pour le type d'activité 105, c'est la copie pour laquelle le plus d'informations sur la conduite défaillante (matériau et/ou diamètre) étaient disponibles qui a été conservée.

2.1.4.4. Gestion des informations caractérisant la défaillance

Dans l'ancienne version de l'affectation des défaillances aux conduites, une tolérance de 1 an était utilisée lors de la comparaison entre la date de casse (DDC) connue et les dates de pose (DDP) des conduites compatibles avec la géolocalisation de la défaillance (dans l'optique d'accepter des imprécisions éventuelles sur les dates). Cette tolérance a été supprimée pour ne pas permettre à des défaillances d'avoir eu lieu (1 an) avant la pose de la conduite, ce qui est incompatible avec les exigences de la modélisation avec Casses (un âge négatif du tronçon à la défaillance étant non toléré).

Deux informations caractérisant chaque défaillance sont essentielles pour la compréhension des phénomènes de pertes en eau et de leur prise en charge par la SWDE (notamment concernant le fonctionnement de la RF) :

- S'agit-il d'un crash : O/N ? (voir glossaire)
- La défaillance a-t-elle été localisée grâce à l'équipe de RF : O/N ?

Ces questions sont notamment essentielles pour la tâche 07 qui s'intéresse aux fuites trouvées par la RF lors de recherches systématiques (pour lesquelles les réponses sont NON et OUI respectivement). Plusieurs sources d'informations dans le SI de la SWDE ont dû être utilisées pour décider de la réponse à ces deux questions pour chaque défaillance (voir Annexe 3).

Pour finir, 8 informations de dates dans les divers fichiers de données utilisés pour le projet sont pertinentes pour la localisation et la réparation des défaillances. Chacune de ces dates a été utilisée selon un ordre de priorité décrit dans le Tableau 3 et caractérisant la fiabilité de la source d'informations. En cas d'absence de la source de données de rang n, la source de rang n+1 a été utilisée. La date de l'OT est toujours connue, sauf pour les points SIG sans OT (voir 2.1.4.2).

Tableau 3 : Dates utilisées pour la localisation et la réparation des défaillances

Priorité	Source de données	Date de localisation	Date de réparation (date de casse)
1	Anomalies ou opérations SIG	DATECONS (anomalies seulement)	DATEREPA

2	Fichiers fuites (Excel)	date de localisation	date de réparation
3	SAP	date de l'OT	date de début réel
4	SAP		date de début planifié
5	SAP (OT archivés)		date de fin de panne
6	SAP		date de l'OT

Il est important de noter que pour DATEREPA, la valeur 31/08/2021 a été ignorée pour les anomalies SIG car elle correspond à une complétion en masse le 31/08/2021 de dates de réparation non renseignées.

Enfin, pour environ 8 400 défaillances, les dates de localisation et de réparation étaient inversées (réparation avant localisation), essentiellement à cause d'imprécisions probables dans la saisie des dates et d'erreurs de frappe facilement détectables (comme 15/06/2007 au lieu de 15/06/2008). Pour ces dernières, les dates ont été remises dans le bon ordre (écart de dates de moins d'une semaine dans plus de 90 % des cas).

2.1.4.5. Gestion des erreurs topologiques du SIG

De nouvelles corrections topologiques ont été apportées aux fichiers de conduites. Le script R développé sur Mons a permis de trouver des dizaines d'erreurs topologiques sur le réseau en service des autres secteurs et de les corriger manuellement (voir Annexe 5). Ce script R a été adapté pour corriger automatiquement plusieurs centaines d'erreurs qui n'auraient pas été détectables et traitables efficacement à la main sur le réseau HS (voir Annexe 5).

2.2. Validation de la méthode d'affectation des défaillances de conduites aux tronçons du SIG

Près de 20 % des réparations de conduite sont pré-affectées aux tronçons du SIG car elles correspondent à des « anomalies SIG » ou à des « opérations SIG » pour lesquelles l'ID de conduite défaillante est fourni dans la table attributive. Pour ces dernières, il est possible de comparer l'ID de conduite connu à celui qu'on obtiendrait sans cette information de géolocalisation précise (donc uniquement avec l'information de rue et éventuellement du numéro d'immeuble), ce qui constitue une validation de la méthode d'affectation des défaillances aux conduites. Cette validation a été effectuée à partir des données de défaillances de conduites (type d'activité 105) sur les secteurs de MAG, soit sur 867 réparations de conduites pour lesquelles cette comparaison est possible. L'Annexe 4 détaille la méthode employée et ses résultats. En résumé, 60 % des affectations comparées sont identiques, 76 % concernent des conduites avec les mêmes caractéristiques (MAT, DIA, DDP), et 89 % sont dans la même rue. Le succès de l'affectation dépend essentiellement de la précision des informations sur la défaillance de conduite (quelle adresse est donnée avec quelle précision et quelles caractéristiques de la conduite réparée sont connues). Ainsi, nous avons décidé d'attacher à chaque défaillance de conduite un degré de fiabilité de son affectation à une conduite (Tableau 4).

Tableau 4 : Indicateur d'incertitude pour chaque défaillance de conduite affectée aux tronçons du SIG.

Informations disponibles		Indice de fiabilité d'attribution	Type d'attribution
Numéro d'immeuble	Données MAT ou DIA		
NON	NON	Mauvaise	Incertaine

NON	OUI	Douteuse	Vraisemblable
OUI	NON	Correcte	
OUI	OUI	Bonne	
SIG		Certaine	

Lors de la prédiction des défaillances de conduites avec le modèle LEYP et de la représentation graphique des résultats sous SIG (voir 4.3.1.3), pour un même risque prédit de défaillance de conduite, une distinction sera faite entre les conduites qui ont vraisemblablement déjà été défaillantes (indice de fiabilité Correcte à Certaine, en trait plein) et les conduites dont on n'est pas sûr qu'elles aient déjà été défaillantes (indice de fiabilité Mauvaise à douteuse, en pointillés).

Lorsqu'un tronçon du SIG recevra plusieurs défaillances de conduites aux indices de fiabilité différents, il conservera l'indice de fiabilité le plus élevé, car un tronçon ayant déjà subi des défaillances possède plus de chances d'en subir à nouveau qu'un tronçon n'en ayant jamais subi.

2.3. Bilan de l'affectation des défaillances de conduites aux tronçons du SIG

2.3.1. Résumé

Le Tableau 5 présente les résultats de l'affectation des défaillances de conduites (activités SAP n°105 et n°999) pour les 17 secteurs en quelques statistiques essentielles. Le

Tableau 6 se penche sur l'influence de l'année et de la méthode sur le nombre d'affectations.

Tableau 5 : Bilan des défaillances de conduites affectées aux tronçons du SIG.

Caractéristique	Valeur		Critère de distinction		
			Conduite ES	Conduite HS	
Nombre de défaillances de conduites	48 834				
			45 097	3 737	
Tronçons agrégés	Nombre	280 086	266 657	13 429	
	km	29 599	28 214	1 385	
Tronçons agrégés concernés par des défaillances de conduites	Nombre	26 625	25 279	1 346	
	km	6 179	5 868	311	
Méthode d'attribution	48 834		SIG	Rue + Numéro	Rue
			10 117	9 607	29 110

Tableau 6 : Détail du nombre d'affectations par méthode et par année.

Année	Affectation par la rue	Rue et numéro d'immeuble	Anomalie ou opération SIG	Total
2011	4 204	152	0	4 356
2012	4 434	164	2	4 600
2013	4 081	107	245	4 433
2014	3 335	85	734	4 154
2015	2 969	419	938	4 326
2016	2 211	1 173	705	4 089
2017	2 390	1 696	490	4 576
2018	2 159	2 170	477	4 806
2019	1 356	1 481	1 588	4 425
2020	1 085	1 108	2 023	4 216
2021	871	1 048	2 915	4 834
Total général	29 110	9 607	10 117	48 834

On observe au fil des années une diminution progressive du nombre défaillances pour lesquelles seule la rue est renseignée au profit de défaillances géolocalisées directement dans le SIG. Cette diminution s'accroît fortement entre 2019 et 2021 par mise en place définitive des anomalies SIG. La nette augmentation du nombre de géolocalisations grâce au numéro d'immeuble à partir de 2016 est justifiée par l'apparition des fichiers « fuites » qui le précisent souvent.

2.3.2. Phénomène de concentration des défaillances de conduites

On constate que seulement près de 10 % des tronçons (21 % du linéaire) ont subi des défaillances de conduite, au rythme de 1.8 fois par tronçon défaillant. De plus, 35 % de ces tronçons défaillants (9 423/26 625, soit 45 % du linéaire défaillant) le sont au moins 2 fois. Ceci traduit un phénomène de concentration des défaillances de conduites sur certains tronçons. Ce phénomène est attendu et justifie d'ailleurs l'utilisation du modèle LEYP pour les prédire. Néanmoins, on peut se demander si le nombre particulièrement élevé d'affectations des défaillances de conduites par la rue (60 % des affectations), utilisant souvent la plus longue conduite de la rue à défaut d'informations spécifiques, pourrait avoir une influence sur la concentration de certaines défaillances. Un des moyens pour atténuer cet effet serait de ne pas systématiquement choisir la conduite la plus longue de la rue par défaut, mais de prendre aléatoirement une des conduites compatibles de la rue avec une probabilité proportionnelle à la longueur de la conduite. Cette méthode alternative, testée brièvement sur les secteurs de MAG, n'a cependant pas montré d'atténuation particulière du phénomène de concentration et a été abandonnée à cause de son caractère non déterministe dans l'affectation des défaillances aux conduites.

2.4. Étude de l'impact de la complétion de données sur les prédictions de défaillances

2.4.1. Contexte et objectifs

Comme précisé dans le rapport intermédiaire de 2021, certains attributs de la couche des conduites HS, comme la DDP, ont dû être complétés par une technique de propagation de l'information existante de proche en proche (moyennant une recherche de l'attribut sur le réseau en service voisin en cas d'absence totale de l'information), afin de minimiser le nombre de valeurs manquantes et de pouvoir exploiter au mieux les défaillances des conduites HS dans le modèle. Bien que cette complétion possède un caractère bénéfique évident, il est en réalité difficile de savoir exactement de quelle façon elle influence les résultats de modélisation et avec quelle importance pour quels paramètres.

Aurélien Mirebeau (étudiant en fin de parcours d'école d'ingénieur) a été recruté au sein d'INRAE pour un stage de 6 mois (mars à août 2022 inclus) pour contribuer au projet GePaME. L'objectif de son stage a ainsi été d'étudier les conséquences de la complétion de données sur les prédictions de défaillances avec Casses (c'est-à-dire le modèle LEYP). L'hypothèse de base, du point de vue de la modélisation des défaillances, est qu'une information déjà bien connue à l'origine est supérieure à une information bien complétée, cette dernière étant supérieure à une absence d'informations (remplacement de l'attribut par la valeur « inconnu »), cette dernière étant supérieure à une information mal complétée (donnée complétée différente de la donnée d'origine). Le rapport de stage est fourni en annexe AX1 séparée du présent rapport. Le résumé suivant rappelle les éléments essentiels de ce travail, dont les détails sont reportés à l'Annexe 6.

2.4.2. Résumé de l'étude

Dans un premier temps, une recherche bibliographique a été faite pour identifier des améliorations possibles de la méthode de propagation de proche en proche de l'attribut manquant des conduites utilisée dans GePaME (Rodriguez et al., 2021). La technique d'origine a ainsi bénéficié de 2 améliorations optionnelles : (1) le conditionnement de la propagation de l'attribut manquant (par exemple la DDP) par la cohérence d'un autre attribut (par exemple le matériau) entre les conduites immédiatement adjacentes ou (2) entre les conduites non immédiatement adjacentes.

Un ensemble de 4 méthodes de complétion comprenant la méthode originelle et ses 2 améliorations ainsi qu'une méthode moins élaborée (estimation par la médiane ou le mode) a été employé dans un test de validation comparant les performances de ces 4 méthodes en termes de taux de complétion (quantité) et de taux d'erreurs de complétion (qualité) de la DDP par rapport à la donnée d'origine (en se limitant au réseau ES qui est bien connu et connexe). Ceci a révélé que les améliorations de la méthode originelle présentent des avantages sur la quantité ou la qualité de la complétion, et que dans l'ensemble la méthode d'origine possède une performance de complétion satisfaisante (en particulier comparée à une estimation « naïve » par la médiane).

L'utilisation de données du réseau de conduites ES, dans leur version d'origine, et dans des versions altérées représentant divers scénarios de manques de l'attribut matériau et de complétion de ces derniers par les 4 méthodes étudiées a permis de confirmer l'hypothèse de base grâce à des calages du modèle LEYP pour chacun des scénarios. Ainsi, les améliorations proposées de la méthode de complétion permettent généralement de se rapprocher un peu plus des données d'origine et des résultats de prédictions des défaillances des tronçons qui seraient obtenues avec ces dernières (bien que la méthode de complétion d'origine ne s'en éloigne que de peu).

2.5. Préparation des données issues de la RF pour la tâche 07

Pour rappel, la tâche 07 a pour objectif de développer un modèle (régression logistique) qui permet de calculer pour chaque tronçon la probabilité qu'il soit dans un état fuyard (affecté par une « fuite », voir glossaire). Il est important de préciser que le modèle ne s'intéresse pas en théorie aux crashes, qui ne demandent pas la même approche que la RF systématique. Cette modélisation demande une grande quantité de données en entrée, dans un format adapté au modèle.

2.5.1. Besoins en données de la tâche 07 et données disponibles

Les données d'entrée concernées sont :

- les tronçons de canalisation ainsi que leurs caractéristiques patrimoniales, géographiques, et d'exploitation ;
- ces mêmes caractéristiques agrégées au niveau des cellules des tronçons ;
- les résultats (fuite trouvée ou non sur quel tronçon) des inspections par la recherche de fuites (RF).

Les données brutes de la SWDE doivent être traitées pour satisfaire ces besoins. Elles correspondent à :

- deux fichiers SIG de conduites (en service et hors service) précisant leur caractéristiques patrimoniales ;
- un fichier Excel de « prestations » ponctuelles de RF sur certaines cellules depuis 2016, indiquant la date d'inspection et le temps passé, mais ne précisant ni la raison de l'inspection, ni la méthode employée, ni les tronçons inspectés, ni le résultat de l'inspection, ni le lien avec les autres prestations des jours précédents ou suivants ;
- un ensemble de fichiers décrivant les réparations de fuites et de crashes de conduites ou d'autres éléments du réseau effectuées depuis 2007 à diverses adresses géolocalisées de façon plus ou moins précise (point SIG ou rue avec/sans numéro d'immeuble).

Les traitements pour générer un fichier de défaillances affectées aux ID de tronçons du SIG ont été détaillés dans le rapport intermédiaire de 2021 ainsi que dans la section 2.1 du présent rapport. Le traitement des caractéristiques des tronçons et de leurs cellules fait l'objet de la tâche 05. Le rapport intermédiaire de 2021 détaille notamment comment des caractéristiques additionnelles (géographiques, d'exploitation) des tronçons ont été ajoutées (voir la section 6.1 du présent rapport pour les mises à jour de certaines méthodes et l'ajout de certains attributs). La construction des caractéristiques agrégées par cellule est détaillée en section 6.4.6.

Les sections suivantes détaillent comment les données propres à la RF sont traitées, en particulier le rattachement des inspections par la RF aux tronçons du SIG et la détermination du résultat de l'inspection.

2.5.2. Nettoyage des fichiers de prestations de RF

La recherche active de fuites (RF), sur une cellule identifiée comme siège potentiel de fuites, procède par la mise en œuvre de prestations impliquant :

- le déploiement sur le terrain d'équipes techniques chargées de la pré-localisation de la portion de cellule sur laquelle la présence de fuites est la plus vraisemblable,
- puis de l'inspection acoustique systématique, tronçon par tronçon, des éléments du réseau de la zone pré-localisée.

Il est à noter que la mise en œuvre de la pré-localisation n'est pas systématique, mais soumise à l'appréciation des équipes en charge de la RF.


93 590 prestations de RF ont été enregistrées dans SAP à la date de transmission du fichier des prestations par la SWDE (08/03/2022). Ce dernier contient cependant parfois les mêmes prestations encodées plus d'une fois pour un même agent de RF, ainsi que des prestations qui ont été annulées avant d'avoir été effectuées (dans ce cas des prestations concomitantes mais à durées négatives sont présentes). Afin de ne conserver que les prestations réellement effectuées et sans doublons, un nettoyage a été appliqué. Ce dernier a consisté à ne conserver qu'une seule copie (obligatoirement à durée positive) parmi les prestations concomitantes et effectuées par un agent donné dont la somme des heures prestées sur l'ensemble des copies est strictement positive. Ce traitement réduit de 7 411 le nombre de prestations.

2.5.3. Regroupement en campagnes de RF

Le plus souvent initiées par l'observation d'une dérive anormale de la consommation d'eau de la cellule, les prestations sont généralement menées sur plusieurs jours ouvrés (plus ou moins) consécutifs, jusqu'à l'identification d'au moins un emplacement fuyard, ou bien l'abandon des investigations si aucun résultat concluant n'est obtenu dans un délai raisonnable. L'ensemble de ces prestations peut s'étaler sur une ou plusieurs semaines et est appelé « campagne de RF ». La notion de « campagne de RF » est absente du SI de SWDE mais nécessaire à la mise en œuvre de la modélisation de la probabilité d'état fuyard des tronçons.

La méthode (heuristique) employée pour définir de telles campagnes, détaillée en Annexe 8, part du principe que des prestations journalières distantes de 3 jours ouvrés ou plus font partie de deux campagnes différentes. Ce délai-seuil optimal de 3 jours a été estimé sur l'ensemble des 17 secteurs grâce à un indicateur prenant en compte le nombre de campagnes obtenues et leur densité. Suivant ce principe de regroupement (voir Figure 1), deux prestations ayant eu lieu consécutivement un jeudi et le mardi de la semaine suivante (+3 jours ouvrés) ne font pas partie de la même campagne, ce qui n'est pas le cas si la deuxième prestation a lieu le lundi et non le mardi de la semaine suivante (+2 jours ouvrés). 33 502 campagnes sont ainsi obtenues et nommées par un identifiant de la forme NumCellule_NumCampagne (par exemple 51108103_1 pour la campagne n°1 sur la cellule 51108103). L'Annexe 8 propose une brève discussion de la pertinence de l'indicateur utilisé pour définir les campagnes et de sa valeur « optimale ».

Jour	L	Ma	Me	J	V	S	D	L	Ma	...
Inspection	X	X		X					X	...
Num. de campagne	1	1	1	1					2	...



 +1 jour +2 jours +3 jours

Figure 1 : Illustration du regroupement en campagnes pour une cellule.

2.5.4. Lien entre campagnes de RF et fuites trouvées

Une fois les campagnes créées, les défaillances (fuites et crashes) localisées par la RF (sur conduite ou autres) ont pu être explicitement liées aux campagnes de RF. Les critères utilisés ont été les suivants :

- La campagne a eu lieu sur la même cellule que la défaillance.
- La campagne a débuté avant la localisation de la défaillance.
- La défaillance a été trouvée par la RF (attribut « par RF » = 1 dans les fichiers fuites, voir 2.1.4.4).

En cas de multiples campagnes satisfaisant ces critères, c'est la campagne la plus récente qui a été choisie. Le fichier des défaillances utilisé pour ce traitement est le fichier traité des défaillances (avec les types d'activités pertinents pour les fuites et crashes, voir 2.1.2) affectées aux tronçons du SIG (98 690 défaillances). Suite à ce rattachement, 19 858 défaillances parmi les 23 920 concernées par une intervention de la RF ont été associées à un ID de campagne RF, ce qui représente 12 491 campagnes avec au moins une défaillance localisée par RF (soit environ 37 % des campagnes de 2016-2021). La section 8.3.1 se penche sur la notion de succès des campagnes RF.

2.5.5. Table d'entrée pour la régression logistique

Le fichier d'entrée pour la régression logistique a été construit de façon à, pour chaque campagne de RF, lister l'ensemble des tronçons de la cellule en service au moment de l'inspection ainsi que toutes leurs covariables, et afficher le résultat de la campagne pour chaque tronçon (fuite ou crash trouvé sur quel élément du réseau porté par le tronçon). De cette façon, il n'est possible d'inclure par exemple dans l'analyse que les seules campagnes qui ont soit localisé au moins une fuite non crash de conduite, soit rien du tout (voir 8.2 où des filtres supplémentaires sont détaillés pour les besoins précis de modélisation). Un compteur du nombre de défaillances subies par chaque tronçon strictement avant chaque inspection par type de défaillance (fuite vs crash) et par type d'objet concerné par la défaillance (conduite vs autre) a été ajouté pour servir de covariable spécifique à la régression logistique. La table d'entrée pour la régression logistique possède, pour la période 2016-2021, plus de 4.5 millions d'individus statistiques, qui sont les tronçons avec un âge donné (un même ID de tronçon inspecté par 2 campagnes constitue 2 individus distincts). L'intégration de ces données dans le modèle est décrite dans la section 8 du présent rapport.

3. Tâche 02 – Analyse de la gouvernance de la SWDE en matière de GPI

3.1. Présentation de la tâche 02

La tâche 2 du projet GePaME porte sur la gouvernance territoriale de la gestion patrimoniale de la SWDE. Cette partie du projet est née de l'hypothèse suivante : la gestion patrimoniale des infrastructures (GPI) de la SWDE ne dépend pas uniquement de décisions et de modes d'organisation internes, mais aussi du fonctionnement général du secteur de l'eau potable en Wallonie, et des relations qu'entretient ou non la SWDE avec d'autres acteurs territoriaux et régionaux.

D'une manière générale, on peut définir la gouvernance comme « *un processus de coordination d'acteurs, de groupes sociaux, d'institutions pour atteindre des buts discutés et définis collectivement* » (Lascoumes, Le Galès, 2018). Pour préciser cette notion, nous parlons plutôt ici de gouvernance territoriale, qui resserre l'étude de ces processus de coordination sur un territoire particulier, avec des enjeux territorialisés et locaux (Rey-Valette et al, 2014). Et en effet, le territoire sur lequel interviennent les acteurs de l'eau wallons n'est pas insignifiant : il s'agit d'un construit social, au sens

d'un espace d'action particulier cadrant l'action collective, un socle de négociation où opèrent des acteurs divers qui se l'approprient différemment (Lardon et al, 2008).

Analyser la gouvernance territoriale de la GPI de la SWDE nous impose donc pour commencer d'identifier les enjeux de ce territoire à l'égard de l'eau potable, de comprendre les institutions et les dispositifs structurant le champ des possibles pour la SWDE, et d'identifier les acteurs importants du secteur de l'eau potable en Wallonie.

3.2. Appropriation du contexte : enjeux et acteurs

La première partie du travail a consisté à contextualiser notre problématique et les enjeux territoriaux, par le biais d'une exploration de la « littérature grise » provenant des acteurs du secteur et du Service Public Wallon (SPW), complétés d'articles académiques.

Ainsi, nous avons pu identifier l'abondance de la ressource en eau sur le territoire wallon, faisant de lui le réservoir de Bruxelles et de la Flandre (Aubin, Varonne, 2001). Toutefois, ces ressources sont inégalement réparties à l'échelle du territoire, et il existe aussi des problèmes locaux de surexploitation (SPW, 2022). La qualité de l'eau utilisée (à 80% des eaux souterraines) est plutôt bonne, même si on observe une hausse constante depuis les années 2000 du nombre de fermetures de points de captage pour cause de pollution, notamment au nitrate (SPW, 2017). Enfin, cette situation est fragilisée par le dérèglement climatique, pouvant conduire à des épisodes d'intense sécheresse (e.g. 2016 à 2019) et/ou de précipitations surabondantes provoquant des inondations (e.g. été 2021). De ce point de vue, l'eau « ressource » constitue un enjeu important en Wallonie, puisqu'il conditionne en partie les capacités de production et de distribution d'eau potable de la SWDE.

Nous nous sommes aussi intéressés au fonctionnement du secteur de l'eau, en termes de système d'acteurs, de distribution des rôles et d'instruments structurants pour la GPI de la SWDE. Nous avons pu voir la place centrale de la SWDE dans la production / distribution, puisqu'elle en gère 70%, les intercommunalités et les communes représentant une part de plus en plus ténue (Aquawal, 2020) du fait de la tendance au regroupement des opérateurs. A ce stade, il nous semblait important de comprendre la relation avec la Région Wallonne, puisque la SWDE est sous contrat de gestion avec elle. Est ainsi indiqué dans le contrat de gestion : « *la Région wallonne consacre la SWDE comme son bras opérationnel essentiel en matière de production et de distribution d'eau* » (Gouvernement Wallon, 2017). Nous voulions alors savoir les effets de cette relation sur la GPI de la SWDE. Nous avons aussi identifié la fédération des opérateurs de l'eau : Aquawal, et nous voulions voir si cette structure était un lieu de construction des enjeux du secteur, et d'élaboration d'une vision partagée sur la gestion de l'eau. Ensuite, la question de la formation du prix de l'eau nous a semblé importante. C'est en effet une variable directement responsable de la hauteur possible des investissements, du fait du principe « Coût-Vérité » en vigueur (équivalent de « l'eau paye l'eau » en France). Enfin, nous avons identifié le Schéma Régional Des Ressources en Eau (SRRE) comme étant un dispositif primordial pour la stratégie d'investissement de la SWDE.

3.3. Problématisation et hypothèse

Ce travail préliminaire nous a amenés à nous poser les questions suivantes. Quelles sont les politiques publiques et les dispositifs régissant le secteur de l'eau potable en Wallonie ? Quelles sont les relations de la SWDE avec les autres acteurs de l'eau ? Comment la GPI de la SWDE incarne-t-elle un processus de gestion de l'eau (potable) territorial et trans-sectoriel ? A partir de la thèse selon laquelle la SWDE

est devenue le chef de file de la gestion de l'eau potable en Wallonie, nous avons formulé les hypothèses suivantes :

- La SWDE, du fait de son poids parmi les producteurs / distributeurs d'eau potable, a un rôle opérationnel majeur dans l'évolution de l'action publique liée à l'eau potable.
- La SWDE a investi les principales arènes sectorielles de pouvoir, lui permettant de rendre compatible les instruments d'action publique structurants (SRRE, coût-tarifification, communication, PoWalCo...) avec ses propres enjeux.
- L'incertitude qui encadre la GPI et plus globalement la gestion de l'eau, impose de maintenir une activité de veille et de coopération inter- et trans-sectorielle pour répondre aux aléas.

3.4. L'enquête par entretien

Cela nous a conduit à réaliser une enquête par entretien (juin 2022) auprès d'une dizaine d'acteurs concernés par les enjeux de l'eau et des infrastructures en Wallonie, représentant : la SWDE ; la CILE (Compagnie Intercommunale Liégeoise des Eaux) ; Aquawal ; la Direction Régionale de la Voirie ; la plateforme de coordination PoWalCo. Ce panel d'acteurs nous a ainsi permis d'interroger : les relations entre opérateurs de service ; les espaces et instruments d'échange et de coordination entre ces derniers ; les visions des enjeux de l'eau ; l'influence des variables d'ordre (macro-)structurel (État, marché, climat/environnement...).

Les entretiens ont tous été enregistrés avec l'accord des personnes, et sous condition d'anonymat. Ils sont actuellement en cours de traitement et feront prochainement l'objet d'une analyse thématique. Celle-ci visera à mettre en récit l'évolution et la structuration progressive de la gestion de l'eau et des infrastructures en Wallonie, afin d'en éclairer le mode de gouvernance.

4. Tâche 03 – Implémentation de Casses

Le logiciel Casses repose sur le modèle LEYP de défaillances répétées (Linearly Extended Yule Process, Le Gat, 2014; Le Gat, 2016). Ce modèle se base sur la théorie des processus stochastiques de comptage (Ross, 1996), et plus spécifiquement sur le processus de Yule (Greenwood and Yule, 1920).

Très brièvement, le modèle LEYP s'intéresse au phénomène de défaillances répétées d'une canalisation qui est modélisée par une fonction de comptage $N(t)$. Cette fonction en escalier vaut 0 à $t=0$ et augmente de 1 à chaque défaillance de la canalisation au cours du temps. Le processus de Yule amène une caractéristique fondamentale au modèle LEYP : il a une mémoire. Cela signifie que l'intensité du processus (i.e. l'espérance de la différentielle $E[dN(t)|N(t-)]$) dépend du nombre de défaillances passées, notée $N(t-)$, en plus des caractéristiques inhérentes aux canalisations comme leur âge, leur matériau, leur diamètre, etc. Les détails sur la théorie des modèles LEYP et LEYP2s sont reportés à l'Annexe 7.

Dans le cadre du modèle LEYP, seules les défaillances de conduites, c'est-à-dire les fuites et les crashes de type d'activité SAP 105 (réparation de conduite) ou 999 (point SIG sans OT) sont considérées. Le nettoyage des données de défaillances ainsi que l'attribution des défaillances aux conduites effectués par la tâche 01 ont permis de produire les deux fichiers nécessaires à l'implémentation du modèle LEYP(2s) (voir le paragraphe 2.1).

Pour rappel, ces deux fichiers sont :

- Un fichier des tronçons qui compile tous les tronçons et leurs caractéristiques (date de pose et de mise hors service, longueur, diamètre, matériau, caractéristiques patrimoniales, d'exploitation, et géographiques). **Il s'agit ici des tronçons agrégés par INRAE.**
- Un fichier des défaillances, qui contient toutes les défaillances affectées à une conduite, la date de casse et la fiabilité de l'affectation.

Remarque 1 : Ces deux fichiers restent identiques que le modèle soit effectué directement dans le logiciel Casses ou via le code R (pour le modèle LEYP).

Remarque 2 : Pour rappel, seul le modèle LEYP est fonctionnel dans le logiciel Casses. Le modèle LEYP2s, qui corrige le biais de survie sélective en prenant en compte les mises hors service, est uniquement accessible via le code R. Afin de garder une homogénéité de traitement entre modèle LEYP et LEYP2s, toutes les analyses suivantes présentées ont été effectuées via le code R, avec les mêmes fichiers d'entrées.

Le modèle LEYP de défaillances répétées (voir Annexe 7 de ce rapport et Annexe 12 du Rapport Intermédiaire de 2021) définit plusieurs fenêtres temporelles :

- Une période de calage (Figure 2).
- Une période de validation (Figure 2).
- Une période de prédiction, qui implique de redéfinir une période de calage spécifique à la prédiction (Figure 3).

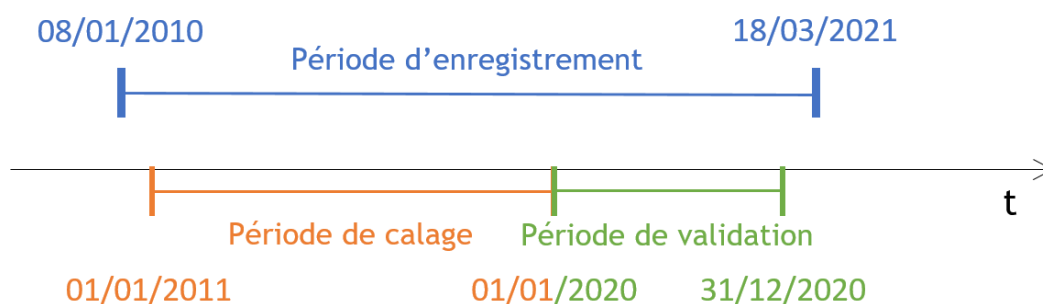


Figure 2 : Représentation des périodes de calage et de validation lors du calage du modèle. Le calage et la validation se font sur des périodes adjacentes comprises dans la fenêtre d'enregistrement totale des défaillances.



Figure 3 : Représentation des périodes de calage et de prédiction pour un objectif de prédiction des défaillances. Lors de la prédiction, le modèle est recalé en utilisant toutes les données disponibles, c'est-à-dire la réunion des périodes de calage et de validation utilisées précédemment.

4.1. Données utilisées

Les calculs avec le LEYP ont été effectués sur les tronçons agrégés des secteurs de MAG. Les calculs avec le LEYP sur l'ensemble des secteurs seront effectués une fois la distribution géographique des défaillances le long des conduites non agrégées terminée (voir section 5.2.1), de façon à avoir les mêmes fichiers d'entrée que la SWDE. Les filtres suivants ont été appliqués sur les fichiers pour le LEYP pour les secteurs de MAG:

- Un filtre pour sélectionner uniquement les tronçons de plus de 6 mètres de longueur.
- Un filtre sur les dates d'observations :
 - Calage : LEYP 01/01/2011-31/12/2019 – LEYP2s : 01/01/2013-31/12/2019
 - Validation : 01/01/2020-31/12/2020
 - Prédiction :
 - « Calage » : LEYP 01/01/2011-31/12/2020 – LEYP2s : 01/01/2013-31/12/2020
 - Prédiction : 2022 et 2022-2026

4.2. Modèles utilisés

Nous avons décidé de caler des modèles différents pour les différents types de matériaux, comme cela se fait de manière usuelle. Nous avons ainsi déterminé trois types de matériaux (et donc calé, validé et prédit avec trois modèles différents) :

- Les fontes (grise + ductile), qui représentent 17 897 tronçons et 3 128 défaillances.
- L'acier, avec 1 902 tronçons et 612 défaillances.
- Les tronçons en AMCI (amiante/asbeste ciment) regroupés avec les autres matériaux, soit 21 915 tronçons et 1 892 défaillances.

Le modèle LEYP(2s) se base sur l'estimation d'un certain nombre de paramètres via un algorithme d'optimisation géométrique (algorithme de Nelder-Mead) :

- α , qui modélise l'effet des défaillances passées.
- δ , qui modélise l'effet de l'âge des tronçons (le vieillissement).
- ψ , qui modélise dans le modèle LEYP2s l'effet des mises hors service contraintes (indépendantes de l'état du tronçon).
- φ , qui modélise dans le modèle LEYP2s l'effet des mises hors service sélective (dues à des défaillances répétées sur le tronçon).
- $\beta_1, \beta_2 \dots \beta_n$, qui modélisent chacun les effets des covariables.

Cependant, selon les données, le modèle n'arrive pas toujours à converger si tous les paramètres sont estimés. Afin de garantir une homogénéité des modèles entre les différents matériaux, nous avons ainsi décidé de fixer certains paramètres afin de garantir la convergence des modèles quel que soit le matériau testé. Nous avons ainsi fixé dans le modèle LEYP δ à 1 (pas d'effet du vieillissement sur l'intensité de défaillances) et dans le modèle LEYP2s δ à 1 et ψ à 2 (effet des mises hors service contraintes proportionnel au temps).

Remarque : fixer δ à 1 ne masque pas nécessairement un effet du vieillissement exploitable pour les prédictions. Même lorsque le modèle converge avec δ non fixé, ses valeurs sont proches de 1 sur les secteurs de MAG. Ceci provient vraisemblablement d'une absence de monotonie, dans les données

brutes, entre le taux de défaillances des conduites (nombre/km/an) et l'âge des tronçons (effet vraisemblable de l'incertitude sur l'année de pose des tronçons agrégés).

Pour chaque modèle de matériau, nous avons testé l'effet de six covariables :

- le logarithme de la longueur du tronçon,
- le diamètre,
- le logarithme naturel du nombre de raccordements +1 (pour éviter un logarithme infini sur les tronçons n'ayant pas de raccordements),
- le pourcentage de la longueur du tronçon en zone très urbanisée (CLC1 : Corine Land Cover 1),
- le pourcentage de la longueur du tronçon sous la chaussée,
- et le pourcentage de la longueur du tronçon sous l'accotement.

Ces six covariables sont communes à tous les modèles de matériaux. Nous avons de plus rajouté une covariable spécifique au modèle des fontes :

- Indicatrice de fonte grise, qui vaut 1 si le tronçon est en fonte grise et 0 s'il est en fonte ductile ;

et une covariable spécifique au modèle AMCI + autres matériaux :

- Indicatrice d'autres matériaux, qui vaut 0 si le tronçon est en AMCI et 1 sinon.

La valeur du paramètre associé à cette indicatrice permet de savoir si, toutes choses égales par ailleurs, la fonte grise casse plus que la fonte ductile (paramètre positif), ou si les autres matériaux cassent plus que l'amiante ciment (paramètre positif).

Nous avons ainsi utilisé six modèles (trois pour le LEYP et trois pour le LEYP2s) avec un calage et une validation, puis les mêmes six modèles pour la prédiction de l'intensité de défaillances dans le futur avec un calage sur toutes les données disponibles (Figure 3).

4.3. Résultats

Le calage fournit une estimation des paramètres des modèles et leur significativité (p -value). Si l'estimation du paramètre est positive, alors l'intensité de défaillances croît avec la valeur de la covariable (e.g. un tuyau plus long aura plus de défaillances qu'un tuyau moins long) et vice-versa pour une estimation négative. La significativité indique si l'effet du paramètre est statistiquement significatif. Si la p -value est inférieure à 0.05, alors l'effet du paramètre sur l'intensité de défaillance est considéré comme significatif (en gras dans les tableaux de résultats).

4.3.1. LEYP

4.3.1.1. Calage

Le calage consiste à estimer la valeur numérique des paramètres à partir des données observées sur la période de calage. Les valeurs des paramètres obtenues sont celles qui donnent les estimations d'intensité de défaillance les plus cohérentes avec les défaillances observées sur l'ensemble des tronçons (formellement, ce sont les valeurs qui maximisent la « vraisemblance » du modèle).

Tableau 7 : Estimation des paramètres du modèle LEYP pour les strates Fontes, Acier et AMCI + autres matériaux. Longueur en mètres. Diamètre en millimètres. Intercept Def : ordonnée à l'origine pour les défaillances.

	Fontes		Acier		AMCI + Autres matériaux	
	Estimation	p-value	Estimation	p-value	Estimation	p-value
α	1.83	0	2.52	0	3.42	0
δ	1	NA	1	NA	1	NA
Intercept Def	-2.48	0	-1.87	0	-3.28	0
Ln(longueur)	0.41	0	0.38	0	0.35	0
Diamètre	-1.03×10^{-3}	2×10^{-4}	-6.92×10^{-4}	0.19	1.00×10^{-3}	1.2×10^{-3}
Ln(nb raccordements + 1)	0.14	0	0.11	0	0.20	0
% CLC type 1	0.45	0	-143	0.97	0.29	6.8×10^{-3}
% sous chaussée	0.19	0.07	-0.45	8.1×10^{-3}	0.54	2×10^{-4}
% sous accotement	0.18	0.09	-0.35	0.03	0.50	4×10^{-4}
Fonte grise	-0.14	0.07				
Autres matériaux					0.03	0.41

On peut remarquer que le modèle estime que la fonte grise casse moins que la fonte ductile (toutes choses égales par ailleurs). C'est un effet probable de la non prise en compte du biais de survie sélective : les fontes grises encore présentes dans les tronçons observés (qui n'ont donc pas été remplacées), sont les plus résistantes de leur catégorie. Ceci masque la faiblesse réelle de ce matériau, traduit par un nombre important de ces conduites qui ont été renouvelées, ce que le LEYP2s peut prendre en compte si les données de renouvellement sont suffisantes.

4.3.1.2. Validation

La validation consiste à comparer les prédictions du nombre de défaillances par tronçon obtenues grâce aux valeurs de paramètres estimées pendant le calage aux défaillances réellement observées sur la période de validation (période n'ayant pas servi à l'estimation des paramètres). Les courbes ci-dessous (Figure 4 à Figure 6) sont la représentation visuelle de cette comparaison. On s'intéresse en particulier aux nombres de défaillances (ou au pourcentage du total observé) qui auraient pu être évitées sur la période de validation en remplaçant par ordre de priorité les tronçons identifiés comme les plus à risque par le modèle (ayant le plus haut taux de défaillance prédit par an et par km). Des indicateurs numériques peuvent aider à juger de la qualité du modèle : aire sous la courbe élevée (« Area »), et pourcentage élevé (e.g., 10 %) des défaillances évitées en changeant un faible pourcentage du réseau (e.g., 1 %). L'axe des abscisses (Figure 4) utilise un % de longueurs de conduites cumulées par rapport à la longueur totale du réseau (qui possède un côté plus pratique), mais il est également possible de le représenter en % du nombre de conduites cumulé par rapport au nombre total du réseau. La courbe bleue utilise les défaillances prédites par le modèle, alors que la courbe rouge utilise les défaillances observées. Ces courbes sont comparées à un remplacement aléatoire des tronçons (première diagonale du cadre).

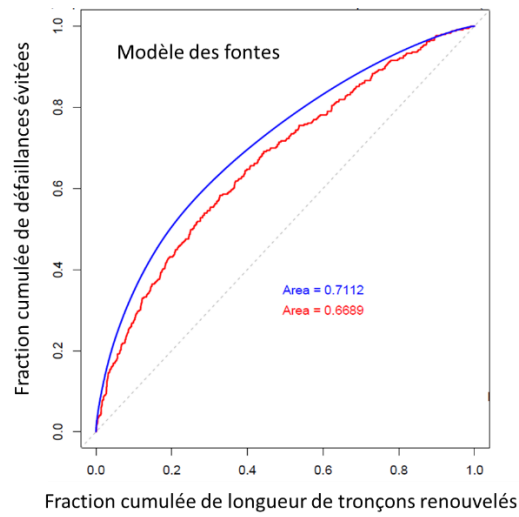


Figure 4 : Courbe de validation des fontes du modèle LEYP.

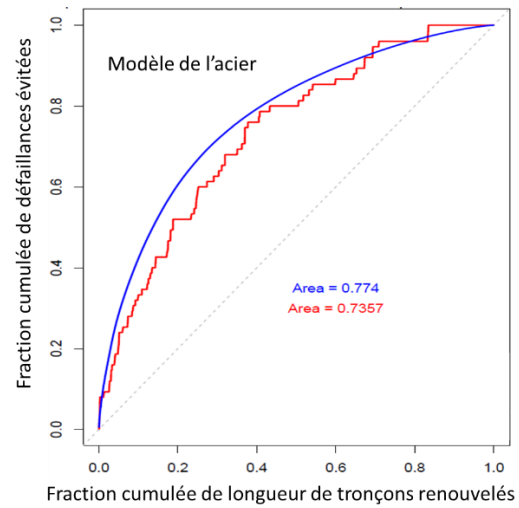


Figure 5 : Courbe de validation de l'acier du modèle LEYP.

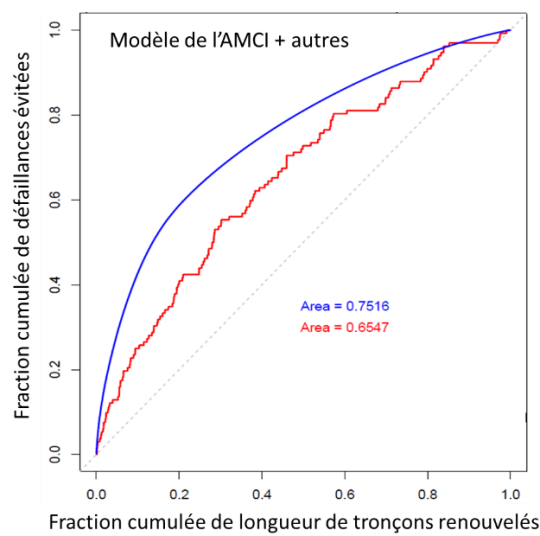


Figure 6 : Courbe de validation de l'AMCI + autres matériaux du modèle LEYP.

4.3.1.3. Prédiction

La prédiction consiste à calculer le nombre de défaillances futures des tronçons grâce aux paramètres estimés sur la période totale (calage + validation), et à ordonner les tronçons par taux de défaillance prédit du plus fort au plus faible. Les longueurs cumulées de conduites en % du total de longueur sont alors calculées pour ces tronçons ordonnés, et utilisées pour créer des catégories de risque de défaillance pour les tronçons. Il est utile de remarquer que les tronçons HS ne sont pas représentés dans les prédictions de risques, bien qu'ayant servi aux calculs.

Remarque : le terme « risque » est utilisé ici dans le sens d'un aléa, et non d'une probabilité multipliée par une conséquence, comme on peut le trouver dans le vocabulaire de la gestion des risques.

Ici, nous avons distingué 3 classes de risque (via des valeurs seuils couramment utilisées) :

- Risque élevé : jusqu'à 2 % (inclus) des longueurs cumulées des tronçons les plus à risque. Les tronçons sont représentés en rouge sur le SIG.
- Risque modéré : entre 2 % (exclus) et 5 % (inclus) des longueurs cumulées des tronçons les plus à risque. Les tronçons sont représentés en orange sur le SIG.
- Risque faible : strictement au-delà de 5 % des longueurs cumulées des tronçons les plus à risque. Les tronçons sont représentés en vert sur le SIG.

Comme décrit dans la section 2.2, la fiabilité de l'affectation des défaillances aux conduites a été utilisée pour créer deux sous-catégories pour les risques élevés et modérés. Les tronçons dont les défaillances ont été affectées de façon incertaine (« Mauvaise » à « Douteuse ») sont donc affichés en pointillés, et les autres en trait plein.

Les résultats des prédictions sur Ath, Mons, et Gaurain sont joints via les cartes (format A0) en annexes séparées AX2 à AX4 de ce rapport pour plus de facilité à manipuler et à visualiser. Un extrait est présenté ci-dessous.



Figure 7 : Extrait de prédiction des défaillances à Tournai.

4.3.2. LEYP2s

4.3.2.1. Calage

Tableau 8 : Estimation des paramètres du modèle LEYP2s pour les strates Fontes, Acier et AMCI + autres matériaux. Longueur en mètres. Diamètre en millimètres. Intercept Def : ordonnée à l'origine pour les défaillances. Intercept HS : ordonnée à l'origine pour les mises hors service.

	Fontes		Acier		AMCI + autres matériaux	
	Estimation	p-value	Estimation	p-value	Estimation	p-value
ψ	2	NA	2	NA	2	NA
φ	0.32	0	0.65	6.9×10^{-3}	0.30	0
α	2.23	0	3.25	0	4.56	0
δ	1	NA	1	NA	1	NA
Intercept Def	-2.22	0	-1.83	0	-3.20	0
Ln(longueur)	0.41	0	0.45	0	0.32	0
Diamètre	-1.48×10^{-3}	0	-1.07×10^{-3}	0.11	1.05×10^{-3}	3.6×10^{-3}
Ln(nb raccordements + 1)	0.14	0	0.16	0	0.20	0
% CLC type 1	0.54	0	-159	0.98	0.29	0.85
% sous chaussée	0.19	0.11	-0.77	2×10^{-4}	0.54	7×10^{-4}
% sous accotement	0.12	0.29	-0.63	2×10^{-3}	0.52	1.1×10^{-3}
Fonte grise	-0.35	0				
Autres matériaux					0.02	0.58
Intercept HS	0.02	0.76	-0.35	0.03	-0.41	2×10^{-4}
Longueur	-0.02	0	-2.42×10^{-3}	0.10	-0.01	0

4.3.2.2. Validation

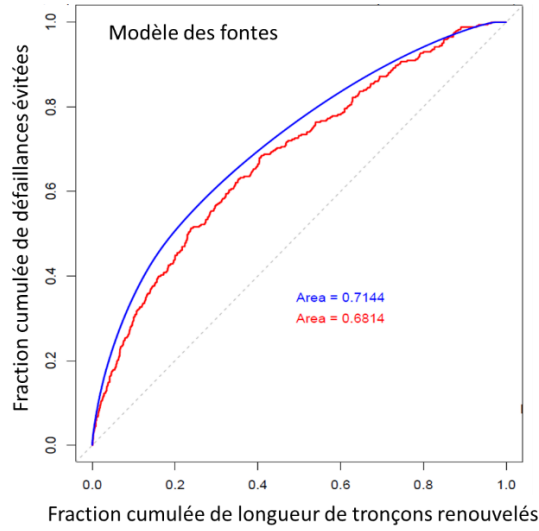


Figure 8 : Courbe de validation des fontes du modèle LEYP2s

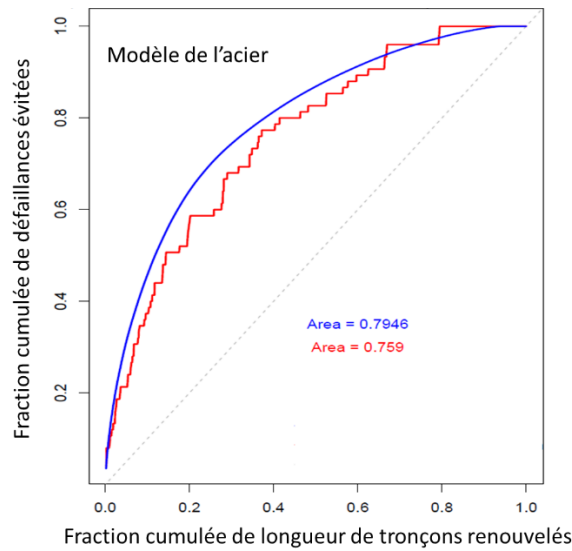


Figure 9 : Courbe de validation de l'acier du modèle LEYP2s

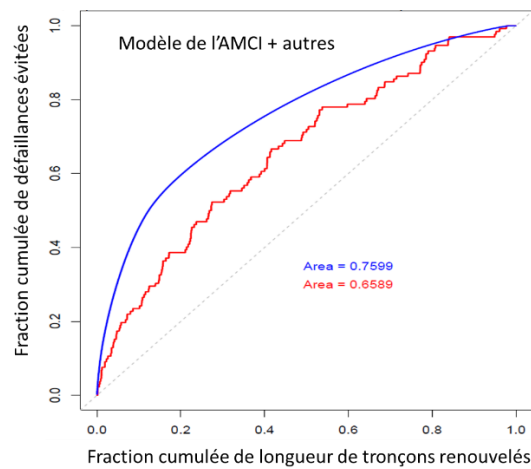


Figure 10 : Courbe de validation de l'AMCI + autres matériaux du modèle LEYP2s

4.3.2.3. Prédiction

Les prédictions avec le LEYP2s et la production de cartes n'ont pas encore été effectuées à ce jour. Elles seront réalisées sur l'ensemble des 17 secteurs.

5. Tâche 04 – Intégration des outils dans le système d'information

5.1. Choix stratégiques

Un groupe de travail restreint SWDE-INRAE (« GT data ») a été constitué pour réfléchir à l'intégration des méthodes de calcul INRAE dans le système d'information (SI) de la SWDE (1^{ère} réunion le 2 mars 2022).

Par soucis de simplification, il a été décidé que les méthodes de traitement des données historiques de la SWDE ne seront pas transférées dans le SI, seul le résultat de ces méthodes sera conservé (tel qu'une archive des défaillances historiques). A contrario, les méthodes d'intégration de données « vivantes » (qui changent dans le temps, comme la position sous chaussée vs accotement) seront implémentées.

5.2. Données historiques

5.2.1. Données sur les défaillances originellement non affectées aux ID de tronçons

Des choix stratégiques ont été faits pour éviter ou simplifier l'intégration des méthodes complexes propres aux défaillances historiques (agrégation des tronçons, affectation des ICAR de rue, affectation des défaillances aux tronçons, ...). Il a été décidé que les défaillances historiques affectées à un tronçon agrégé par INRAE seront rattachées au tronçon le plus long de l'agrégat (revenant ainsi à une version non agrégée des conduites, conforme à celle utilisée quotidiennement par la SWDE). Ce choix permet d'éviter d'implémenter et de rejouer régulièrement la méthode d'agrégation des tronçons. Pour un même type d'activité, quand un tronçon comportera une seule défaillance historique, elle sera localisée en son milieu. Lorsqu'il en comportera plusieurs, elles seront réparties aux positions divisant le tronçon en segments de longueurs égales. Après affectation, les défaillances historiques deviendront des anomalies SIG, et elles porteront des attributs qui les distingueront des anomalies préexistantes et futures et préciseront la fiabilité de leur affectation. Il est important de noter que ce choix suppose que progressivement, et de façon exhaustive avant le 31/12/2022, toutes les nouvelles défaillances (fuites et crashes, sur tous les éléments du réseau) soient intégrées comme des « anomalies » dans le SIG et directement rattachées aux tronçons par la SWDE.

5.2.2. Données complétées sur les tronçons Hors Service

Pour rappel, certains attributs comme la DDP des tronçons HS du SIG étaient parfois manquantes, ce qui a été résolu grâce à une méthode de complétion de proche en proche (voir rapport 2021). Ces données complétées seront transférées à la SWDE afin d'éviter d'avoir à intégrer les méthodes de complétion dans le SI.

5.2.3. Renouvellements de canalisations

Sur la période 2017 – 2021, une méthode automatisée a été développée pour déterminer le motif de pose de nouvelles conduites au sein de chaque cellule : renouvellement ou extension-renforcement.

Les renouvellements de conduites sont distingués par leur proximité avec le réseau HS. Cette proximité s'entend par :

- l'appartenance d'une partie de la conduite ES à un tampon de 20 mètres autour de la conduite HS ;
- la date de pose doit correspondre à la date de mise hors service à un an près ;
- l'angle entre conduite HS et ES doit être inférieur à 30 ° ;
- l'écart de diamètre entre conduite HS et ES doit être inférieur à 200 mm.

Toute autre conduite dont la pose n'est pas liée à une conduite HS est considérée comme une conduite d'extension ou de renforcement (cf Figure 11).

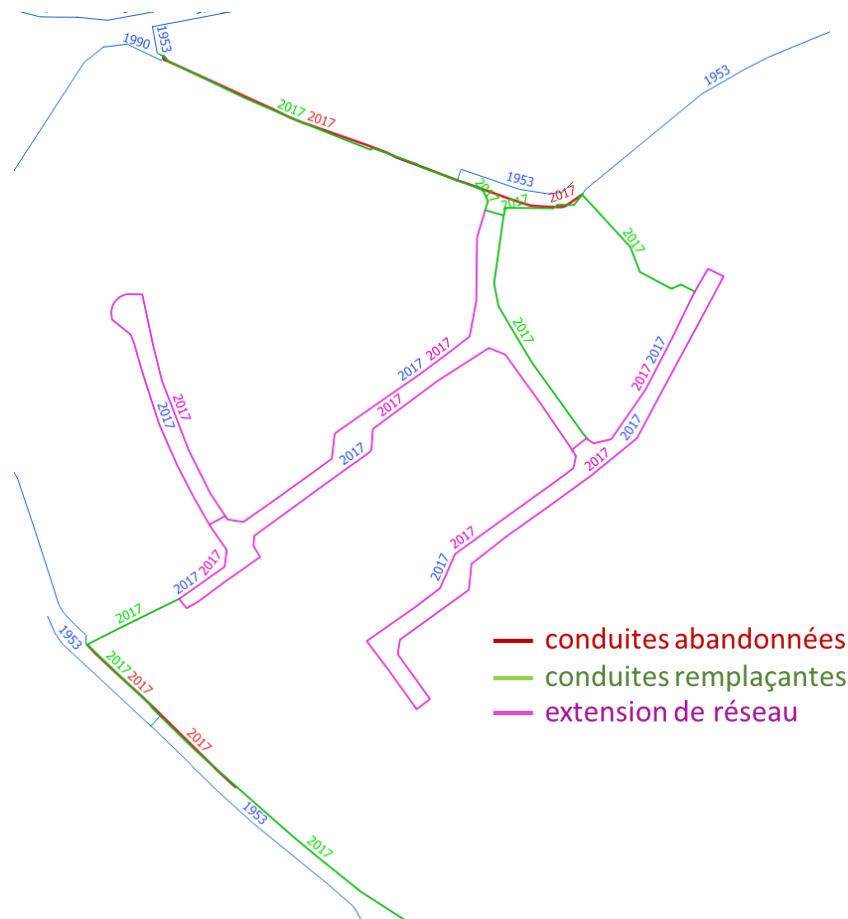


Figure 11 : Renouvellements et extensions du réseau.

5.3. Données vivantes

5.3.1. Covariables additionnelles des tronçons

Les méthodes permettant de calculer des caractéristiques géographiques et d'exploitation attachées aux tronçons seront intégrées pour garantir leur pérennité suivant leur évolution ou celle du réseau :

- Corine Land Cover (CLC) ;
- Type de sol ;
- Utilisation du sol ;
- Corrosivité de l'eau ;
- Pression.

La SWDE a commencé à tester et à valider la retranscription de ces méthodes directement dans FME pour les covariables des tronçons ne nécessitant que des interconnexions de couches SIG (CLC, sols, ...). Le script R de calcul des pressions, plus complexe, est en cours de test pour son intégration via Rcaller, qui permet à FME d'exécuter des scripts R.

La SWDE définit actuellement les raccordements en tant qu'objets dans son SIG. Selon la disponibilité de ces données à la fin de l'année 2022, le calcul du nombre de raccordements (voire d'abonnés) par conduite pourra être effectué soit par la méthode INRAE (via les couches de bâti, voir rapport intermédiaire 2021) si les données sont incomplètes, soit par calcul direct dans le SIG (avec FME par exemple).

5.3.2. Fichier des casses

Le fichier des casses en entrée du logiciel Casses sera à construire au sein du SI de la SWDE à partir des fichiers « anomalies SIG » (historiques + récentes) et « opérations SIG ».

6. Tâche 05 – Exploitation des données de la sectorisation

6.1. Caractéristiques des cellules et covariables des tronçons

6.1.1. Mise à jour du calcul des pressions

La méthode décrite dans le rapport de 2021 a été améliorée par la prise en compte des cotes de trop plein (TP) des réservoirs, qui permettent d'affiner certaines estimations de pression. L'Annexe 9 détaille la méthode mise à jour.

6.1.2. Corrosivité de l'eau

La corrosivité de l'eau distribuée peut être un facteur de dégradation des canalisations de type acier ou fonte non revêtue. L'abaque utilisé par la SWDE (cf. Figure 12) pour caractériser l'agressivité de l'eau distribuée dans chacune de 290 ZDE a permis de définir l'indicateur de corrosivité détaillé dans le Tableau 9 pour chaque tronçon et sa cellule.

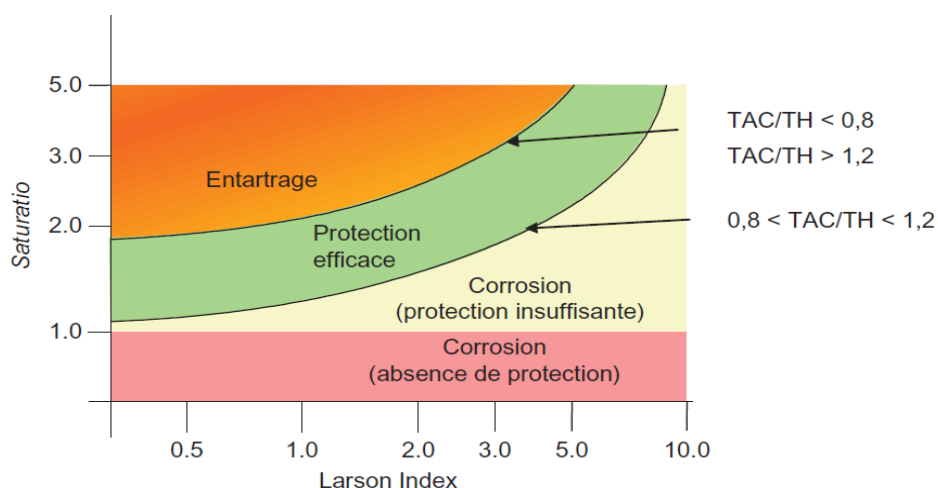


Figure 12 : Caractérisation du facteur agressif de l'eau distribuée par la SWDE (fichier SWDE : Equilibre calco 2022 Complet.xlsx)

Tableau 9 : Indicateur de corrosivité

Indicateur de corrosivité	Qualité de l'eau	Correspondance avec l'abaque
-1	Incrustante	« entartage »
0	RAS	« protection efficace »
1	Agressivité potentielle	« corrosion (protection insuffisante) »
2	Agressive	« corrosion (absence de protection) »

6.1.3. Autres covariables

Nous avons également ajouté les attributs secteur et Zone Fonctionnelle (via la cellule), et le type de voirie (provenant de la couche utilisée pour affecter les ICAR) à chaque tronçon comme nouvelles covariables.

6.2. Stabilité des cellules dans le temps

6.2.1. Problématique

Pour chaque cellule, on dispose de données de débit potentiellement exploitables sur plusieurs années. Par contre, la configuration des compteurs d'entrée et de sortie composant la cellule n'est connue qu'en 2021. L'objectif a été de chercher à comprendre l'évolution de chaque cellule pour valider l'analyse des débits d'une cellule sur l'historique de ses compteurs la plus longue possible.

6.2.2. Tronçons et cellules

L'historique d'appartenance à une cellule de chaque tronçon a pu être construite à partir des tables attributaires (archives Excel) des conduites issues des SIG de 12/2017, 12/2018, 12/2019, 02/2021 et 02/2022. A partir de ces informations, la configuration de chaque cellule a été recomposée chaque année entre 2017 et 2021. La longueur de chaque cellule a pu alors être calculée sur ces 5 années.

Seules les cellules de distribution ont été conservées, et les 103 cellules de moins de 1 500 m ont été écartées. 2 256 cellules restantes ont été analysées.

La structure de la cellule en 2021 a été prise en référence pour détecter les évolutions des années précédentes. Si elles ont été trop substantielles, les débits en seront trop impactés, et l'année n'a pas été conservée.

6.2.3. Démarche adoptée pour la validation de l'historique d'une cellule

La structure de la cellule a été examinée en 3 étapes successives :

1 – Chaque année, la longueur de la cellule a été comparée à celle de l'année 2021 : 2020 a été comparée à 2021, puis si validation, 2019 à 2021, 2018 à 2021, et 2017 à 2021. A chaque fois, les conditions suivantes ont été vérifiées :

- L'écart sur la longueur totale de la cellule doit être inférieur à 10% et à 1 000 m.
- Pour les grandes cellules (longueur de canalisation > 20 000 m), seule la limite de 10 % ci-dessus a été considérée.

2 - Si $n > 2$ années ont été validées (par exemple $n=5$), il a fallu vérifier que les tronçons de la cellule en 2021 sont bien les mêmes qu'en l'année 2021 moins $n-1$ (dans cet exemple 2017). Il faut en effet éviter

les échanges de tronçons entre cellules, tout en prenant en compte les renouvellements ou les découpages de tronçons. La longueur des tronçons en commun entre l'année 2021 moins n-1 (dans cet exemple 2017) et 2021 a été comparée avec la longueur de la cellule de l'année considérée. Si l'écart a été trouvé supérieur à 20 %, cela a été considéré douteux, mais la cellule n'a pas été écartée.

3 – Chaque année, les apports de tronçons en provenance d'autres cellules ont été cumulés en longueur et comparés à la longueur de la cellule en 2021. Les années ont été validées sous conditions variables (Tableau 10). Par exemple, sur 5 ans, l'apport maximum cumulé doit être inférieur à 12 %, ou 5% si la cellule a été classée douteuse au point précédent. Le tableau suivant détaille le nombre de cellules retenues et la période pendant laquelle leur stabilité garantit la validité de l'utilisation des données de débit.

Tableau 10 : Nombre de cellules suivant leur période de stabilité

Nombre de cellules retenues	Période de validité	Apport maximum d'autres cellules	Apport maximum d'autres cellules si douteux
1 613	5 ans de 2017 à 2021	≤ 12%	≤ 5%
173	4 ans de 2018 à 2021	≤ 10%	≤ 5%
139	3 ans de 2019 à 2021	≤ 7%	≤ 5%
179	2 ans de 2020 à 2021	≤ 5%	
152	1 an 2021		
2 256			

6.2.4. Regroupements de cellules

En complément, pour les cellules dont la structure a évolué au-delà des limites fixées, nous avons cherché à savoir si cette évolution résulte d'une division de cellule. Dans ce cas, il est possible de reconstituer les débits à partir des compteurs délimitant la cellule mère (année n) qui sera constituée de plusieurs cellules filles (année n+1). Les 67 groupements de cellules ainsi constitués permettent un suivi des débits sur une période la plus longue possible.

Tableau 11 : Groupements de cellule et leur période de stabilité

Nombre de groupements de cellules retenues	Période de validité
47	5 ans de 2017 à 2021
11	4 ans de 2018 à 2021
7	3 ans de 2019 à 2021
2	2 ans de 2020 à 2021
67	

6.3. Calcul du débit des cellules

Les débits des cellules ont été reconstitués à partir des compteurs d'entrée et sortie de la cellule sur la période de validité précédemment déterminée. Sur les 2 323 cellules ou groupements de cellules, 2 085 ont des données de débit exploitables.

Les Qnuit, Volume jour sont calculés chaque jour (voir rapport 2021), ainsi que l'ILP (Indice Linéaire de Pertes) :

$$ILP_j = \frac{24 \times Q_{nuit_j}}{L}$$

- ILP_j Indice Linéaire de Pertes du jour j en m³/km/j
- L Longueur des canalisations de la cellule en km
- Q_{nuit_j} minimum de la moyenne horaire glissante des débits mesurés par pas de temps de 15 minutes de 01h00 à 05h00 du jour j en m³/h

Le calcul des débits n'a été réalisé que sous certaines conditions :

- Tous les compteurs de la cellule doivent être renseignés.
- Si un débit de la cellule est négatif ou nul, la journée a été supprimée (0h à 23h45).

6.4. Caractérisation des cellules de distribution

6.4.1. Linéaire des cellules

Les cellules étudiées comptent en moyenne moins de 12 km de canalisation, et 11 comptent plus de 50 km.

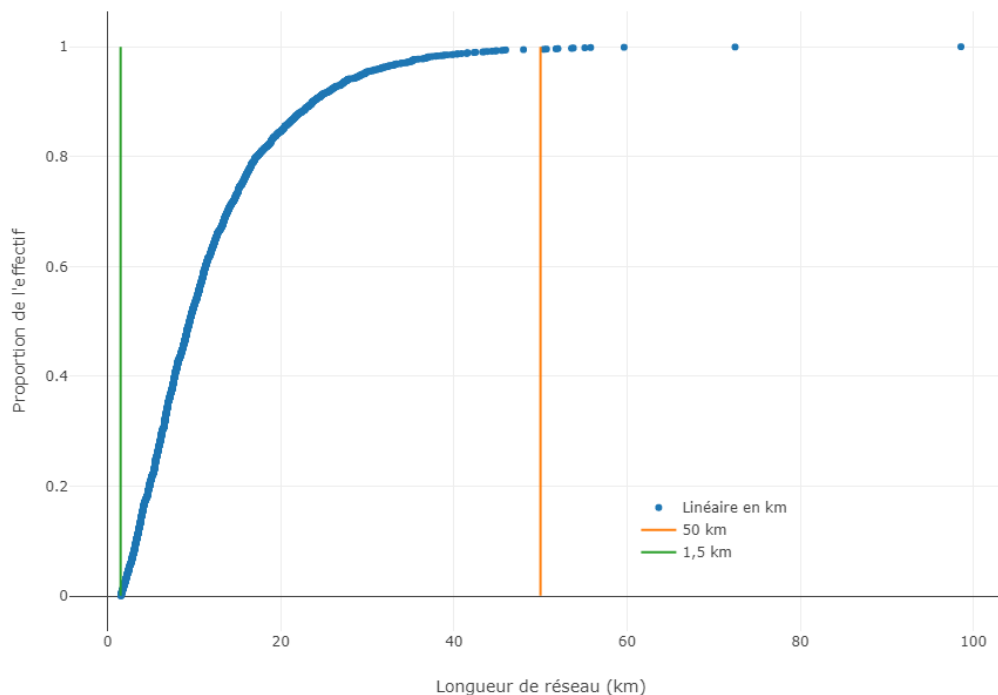


Figure 13 : Répartition des cellules en fonction de la longueur de réseau

6.4.2. Nombre de branchements

Rappel : le nombre de branchements est distinct du nombre de raccordements utilisé par la SWDE. Il s'agit ici du résultat du croisement des couches SIG : point d'adresse (ICAR), bâti (PICC) et réseau AEP (canalisation de distribution et d'adduction mixte).

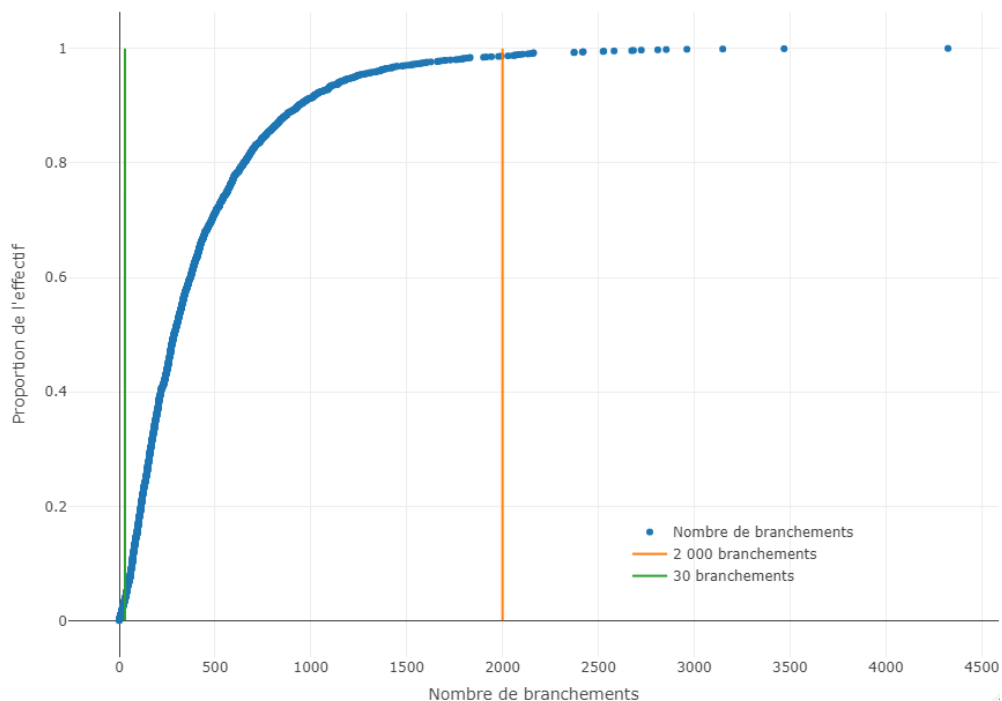


Figure 14 : Répartition des cellules en fonction du nombre de branchements

6.4.3. Densité de branchements

Le nombre de logements et de branchements est lié (cf Figure 15). Un branchement correspond physiquement au raccordement sur la conduite, il peut desservir plusieurs logements (en moyenne 1,15 logement par branchement), le logement pouvant être rapproché de l'abonné.

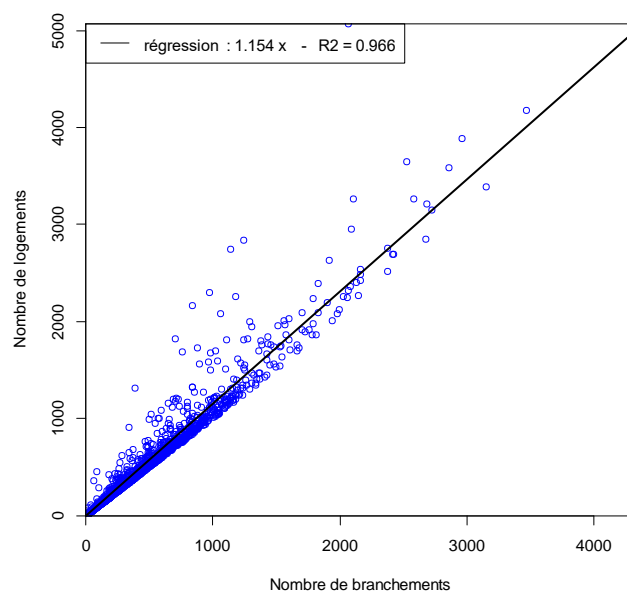


Figure 15 : Nombre de logements par branchement

La densité de branchement ou de logement est liée au caractère rural ou urbain des cellules.

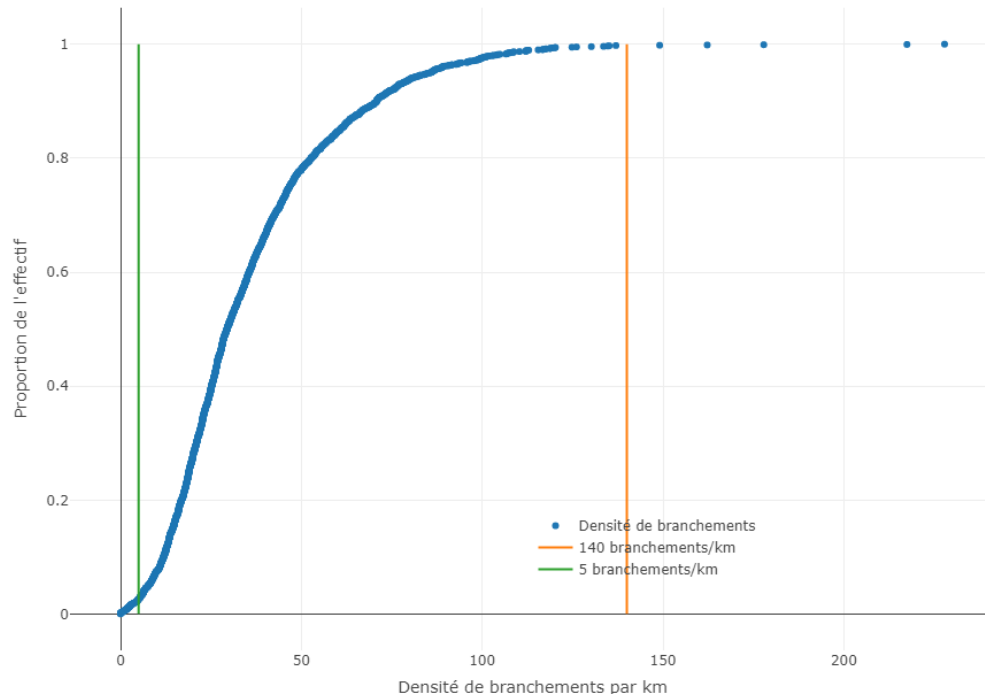


Figure 16 : Répartition des cellules par densité de branchement

6.4.4. Nombre de compteurs par cellule ou groupement de cellule

78 % des cellules sont équipées de 2 compteurs au maximum, mais, exception notable, la cellule 58100108 ne compte pas moins de 19 compteurs.

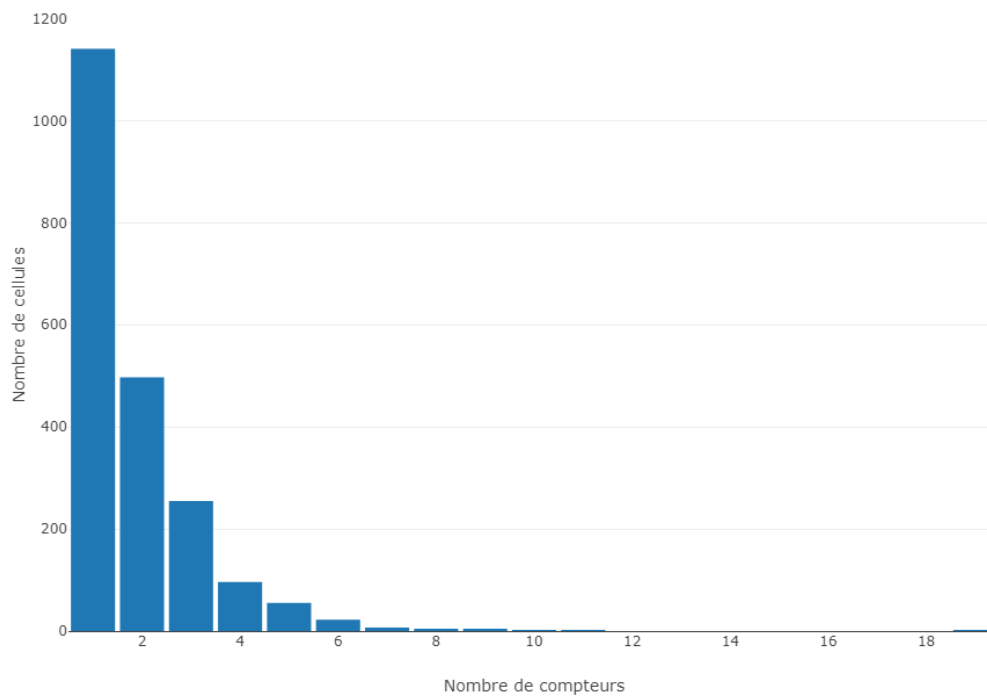


Figure 17 : Nombre de compteurs par cellule

6.4.5. Volume journalier

Le Volume journalier a pu être calculé pendant plus d'un an sur 87 % des cellules et plus de 3 ans pour 46 %.

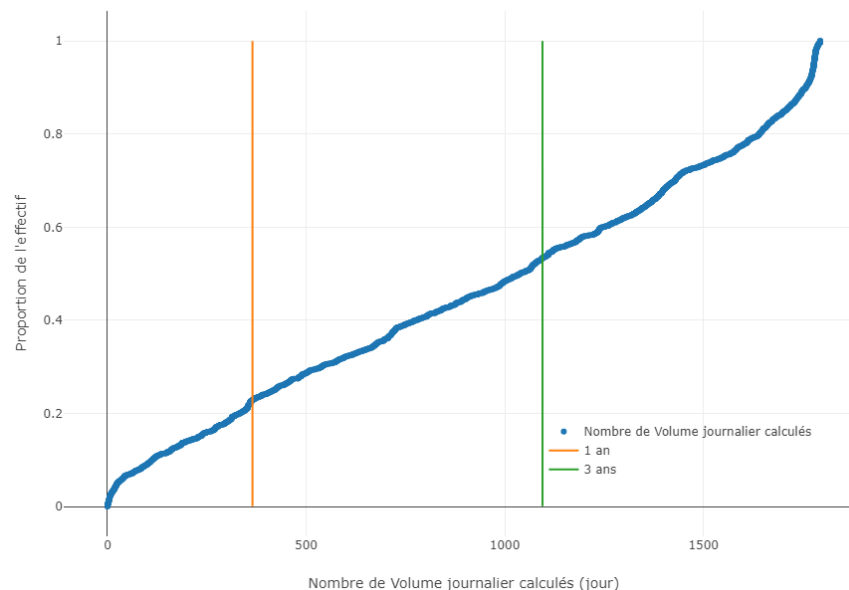


Figure 18 : Répartition des cellules en fonction du nombre de débits journaliers calculés

6.4.6. Typologie des cellules

La construction de la typologie des cellules est en cours. Cela consiste à agréger les données à l'échelle des tronçons, par cellule. Les données concernées sont donc les covariables des tronçons :

- Variables patrimoniales : linéaire de conduites, nombre de raccords, âge moyen, diamètre moyen, pourcentage d'adduction/distribution des conduites, etc.
- Variables d'exploitation : pression, agressivité de l'eau (ZDE), nombre de réducteurs, nombre d'hydrants, etc. Seront notamment inclus les indicateurs de performance tels que l'ILP, lorsque l'ensemble des débits des cellules aura été traité.
- Variables géographiques : couverture, type, et utilisation du sol, hauteur du bâti, etc.

7. Tâche 06 – Analyse des pratiques de recherche de fuites

La tâche 06 a pour objet de décrire et d'analyser le milieu sociotechnique impliqué dans la recherche de fuites pour mieux comprendre les conditions empiriques de mise en œuvre de cette démarche et son intégration dans la gestion quotidienne du réseau d'eau potable wallon.

La description proposée concerne prioritairement les dimensions techniques, sociales et organisationnelles de la recherche de fuites, à savoir :

- Les objets et dispositifs techniques dont l'agencement constitue le système de recherche de fuites (appareillages de détection, réseaux de distribution, bases de données, etc.) ;
- Le(s) dispositif(s) organisationnels développé(s) pour assurer les activités liées à la recherche de fuites (acteurs engagés dans ces activités, division des tâches, protocoles en vigueur et leur application, etc.) ;

➤ Les pratiques et les usages existants.

L'analyse doit permettre de mieux saisir les relations individus/instruments/formes d'organisation et d'évaluer le rôle des interactions qui s'établissent entre ces différentes entités dans la stabilisation et l'appropriation de la recherche de fuites aux activités d'exploitation et de gestion patrimoniale du réseau. L'appropriation renvoie ici à un rapport aux objets et dispositifs et elle relève de régimes organisationnels et juridiques.

La réalisation de la tâche 06 a débuté par un travail documentaire (septembre-décembre 2021) sur la conduite des stratégies de lutte contre les pertes en eau par fuites sur les réseaux de distribution : actions, outils, méthodes et différentes approches utilisées. Ce travail auquel s'ajoutent les informations recueillies lors des premières réunions (réunions d'équipe et CSO) sur les pratiques de recherche de fuites de la part des praticiens et des chercheurs participants, ont permis l'élaboration d'une note méthodologique (cf. Annexe 10). Ce document présente la problématique et les premières hypothèses de la recherche, et précise le protocole d'enquête choisi. Celui-ci comprend des observations directes des activités de recherche de fuites et des entretiens auprès des acteurs impliqués dans ces activités. Ces observations et ces entretiens ont eu lieu entre janvier et mai 2022. Au total, 15 entretiens et 46 séances d'observation ont été réalisés (cf. Tableau 12 et Tableau 13 sur le récapitulatif de l'enquête). Le corpus est en cours de traitement et d'analyse. La rédaction du rapport final de la Tâche 6 est programmé à l'automne 2022.

Tableau 12 : Liste des acteurs interviewés (janvier-mai 2022)

Entretien n°	Fonction des personnes interviewées	Date, lieu
1	Directeur du service « Performance, SIG et RF »	17/01/2022, Loyers
2	Chef de secteur RF	17/01/2022, Loyers
3	Responsable RF (ingénieur)	17/01/2022, Loyers
4	Responsable Distribution-Zone Est (ingénieur)	18/01/2022, Loyers
5	Directeur Pôle Distribution	18/01/2022, Loyers
6	Chef de secteur RF	17/03/2022, Frameries
7	Contremaître Cellule RF	17/03/2022, Frameries
8	Contremaître Cellule RF	25/03/2022, Suarlée
9	Contremaître Cellule RF	25/03/2022, Suarlée
10	Chef de secteur RF	29/03/2022, Couillet
11	Contremaître Cellule RF	31/03/2022, Couillet
12	Chef de secteur RF	11/05/2022, Stembert
13	Contremaître Cellule RF	11/05/2022, Stembert

14	Responsable SIG (ingénieur)	13/05/2022, Villers-Le-Bouillet
15	Technicien Fontainerie	13/05/2022, Villers-Le-Bouillet

Tableau 13 : Liste des observations réalisées (mars-mai 2022)

Secteur RF	Centre d'exploitation enquêté	Nombre des séances d'observation
Mons-Frameries	Mons	13 séances (du 14/03/22 au 22/04/2022)
	Frameries	
Namur-Jodoigne	Suarlée	11 séances (du 23/03/2022 au 31/03/2022)
	Jodoigne	
Charleroi-Couillet	Couillet	10 séances (du 01/04/2022 au 08/04/2022)
	Soignies	
Verviers	Stembert	12 séances (du 09/05/2022 au 19/05/2022)
	Eupen	

8. Tâche 07 - Modélisation de la probabilité d'un tronçon d'être fuyard

8.1. Principe de la modélisation

La maîtrise des pertes en eaux suppose une compréhension du phénomène de fuites des canalisations d'eau potable. Le repérage via les équipes de recherche de fuite et la réparation rapide des fuites limitent les pertes en eau du réseau. Cependant, une connaissance *a priori* de l'état fuyard du réseau permettrait des actions de prévention ou de recherche de fuite ciblée et optimiserait les plans de renouvellement. La tâche 07 de modélisation de la probabilité d'état fuyard des canalisations s'inscrit donc dans cette volonté de meilleure compréhension du phénomène de fuite et de sa durée avant qu'elle n'évolue en crash (voir l'Annexe 12 du rapport intermédiaire 2021).

Cette modélisation se base sur les données des équipes de recherche de fuite, à la fois sur l'action de recherche sur le terrain (une « campagne » de recherche de fuite, ou « campagne RF ») et sur le résultat de cette recherche (une fuite et/ou un crash trouvé(e) ou non à l'issue de cette campagne). La variable d'intérêt est donc le résultat d'une inspection pour chaque tronçon de la cellule inspectée : soit 0 si le tronçon ne fuyait pas, soit 1 si une fuite est détectée sur le tronçon (la valeur 1 est donc donnée uniquement aux fuites –et non aux crashes– ayant eu lieu sur la canalisation elle-même -types d'activités SAP 100, 105 ou 999- trouvée par les équipes de RF). Pour rappel, les données employées pour le LEYP n'incluent pas le type d'activité 100, qui concerne les renouvellements de conduite. Ceci

est fait afin d'écartier les éventuels renouvellements dits « contraints » ou « d'opportunité ». La modélisation par régression logistique peut cependant prendre ce type d'activité en compte du fait du lien explicite du chantier avec le résultat d'une campagne RF, impliquant que le renouvellement en question a nécessairement pour origine une perte en eau identifiée par la RF.

De ce fait, nous étudions ici la probabilité conditionnelle qu'un tronçon soit dans un état fuyard ($Y=1$) sachant qu'il a été inspecté ($D=1$), et selon ses propres covariables (Z) et celles qui ont mené à l'inspection de la cellule (X) : $P(Y = 1 | D = 1, Z, X)$.

La nature binaire (0 ou 1) de notre variable d'intérêt Y nous empêche ainsi d'utiliser des méthodes de régression linéaire classiques qui travaillent sur des variables continues sur $] -\infty; +\infty[$. La modélisation se base ainsi sur une extension de ces modèles de régression, les modèles de régression linéaire généralisée (Generalized Linear Models ou GLM), et en particulier le modèle de régression logistique. De plus, les régressions (logistiques ou non) se basent sur des données censées être indépendantes les unes des autres. Or, ce n'est pas vrai dans le cas présent :

- Un tronçon peut apparaître plusieurs fois dans le jeu de données si sa cellule a été inspectée plusieurs fois ;
- La gestion des équipes de recherche de fuites au niveau des secteurs de distribution implique que tous les tronçons de la SWDE ne sont pas inspectés aléatoirement par une équipe de RF aléatoire mais que les équipes sont réparties par zones géographiques (secteur de distribution, cellules...), ce qui peut avoir un impact sur l'efficacité de la recherche (avec en particulier un « effet secteur »).

Pour pallier ce problème de non-indépendance des données, nous avons comparé la modélisation de la probabilité d'état fuyard avec une régression logistique « classique » (modèles dit fixes, i.e. à « effets fixes ») avec des modèles de régression logistique dit « mixtes », qui contiennent à la fois des covariables fixes et une ou plusieurs covariables dites aléatoires qui permettent de définir des « groupes » de données indépendants les uns des autres. Dans notre cas, les variables aléatoires peuvent être l'identifiant des tronçons, le numéro de cellule ou encore le numéro de secteur de distribution, que l'on peut utiliser pour regrouper des données (voir Annexe 11 pour plus de détails sur la régression logistique et la différence entre modèle fixe et modèle mixte).

8.2. Données d'entrée de la régression logistique (GLM) de l'état fuyard

Le fichier brut produit pour l'analyse de la probabilité d'état fuyard a été détaillé au paragraphe 2.5.5. Pour rappel, il consiste à relier les campagnes RF aux défaillances (crash et fuite) réparées. Une fois ce lien fait, chaque tronçon d'une cellule particulière est considéré comme ayant été inspecté lors d'une campagne pour cette cellule (la liste des tronçons réellement inspectés lors d'une campagne RF n'est pas connue, et il est ainsi postulé que les chances de trouver une fuite sur tronçon non inspecté sont quasi-nulles ; pour une discussion sur cette hypothèse voir le paragraphe 8.4.1). Ceci implique que pour chaque campagne, tous les tronçons de la cellule se voient affecter une indicatrice d'état fuyard qui vaut 1 si une fuite de conduite (type d'activité 100, 105 ou 999) a été détectée sur ce tronçon lors de cette campagne par la RF et 0 sinon. De plus, pour chaque campagne, des variables indicatrices du succès de la campagne sont calculées. Ainsi est reporté pour chaque campagne si l'équipe de RF a trouvé :

- au moins une fuite sur conduite (donc la variable état fuyard vaudra 1),

- au moins une fuite sur un autre composant parmi raccordement, hydrant, appareil (types d'activité SAP 11, 13, 82, 92, 113, et 114),
- au moins un crash sur conduite,
- au moins un crash sur un autre composant.

De plus, sont ajoutées des covariables à ce jeu de données :

- le logarithme de la longueur du tronçon
- le logarithme naturel du nombre de branchements plus 1 (pour éviter les logarithme infini si un tronçon n'a pas de branchement)
- une indicatrice de fonte grise qui vaut 1 si le tronçon est en fonte grise, 0 si le tronçon est en fonte « non-grise » (*i.e.* ductile ou centriflex), et NA sinon
- une indicatrice de pvc qui vaut 1 si le tronçon est en PVC, 0 s'il est en PE et NA sinon
- l'âge du tronçon au moment de l'inspection
- le nombre de défaillances de conduite juste avant l'inspection, qui est la somme du nombre de crashes et de fuites sur le corps de la conduite (type 100, 105 et 999) rapportés entre le début de la fenêtre d'observation (2011) et la veille de la campagne d'inspection RF.

Enfin, les matériaux existants sont regroupés en cinq familles :

- Les fontes : fonte grise + fonte ductile + centriflex.
- L'acier.
- L'amiante ciment.
- Les plastiques : PVC + PE.
- Les autres matériaux : Inconnu + Plomb + Béton Âme Tôle + Polyesther + Inox.

Il faut bien noter que deux analyses ont été menées sur des jeux de données légèrement différents :

1. Une analyse du succès des campagnes de RF sur toutes les données mentionnées ci-dessus (paragraphe 8.3.1) ;
2. Une analyse de la probabilité d'état fuyard où les campagnes RF ayant trouvé uniquement un (ou plusieurs) crash(s) ont été ignorées (paragraphe 8.3.2). En effet, l'hypothèse est avancée que si une équipe RF trouve uniquement un crash, c'est qu'elle a été appelée en renfort par la première ligne pour le localiser, et donc que le crash est déjà pré-localisé plus ou moins finement. De ce fait, l'hypothèse d'inspection de tous les tronçons de la cellule lors de ces interventions est vraisemblablement fautive et ces campagnes sont donc ignorées.

8.3. Résultats

8.3.1. Succès des campagnes RF

8.3.1.1. Etat des lieux des campagnes de recherche de fuite

33 502 campagnes de recherche de fuite ont été effectuées sur le territoire de la SWDE entre le 01/01/2016 et le 31/12/2021. En moyenne, 1 850 campagnes ont été effectuées par secteur de distribution sur cette période, avec un minimum de 705 campagnes à Mons (secteur 5230) et un maximum de 4 718 à Couillet (secteur 5440) (Figure 19, médiane=1 478). Les 2 177 cellules inspectées (sur un total d'environ 2 300, soit 93 %) l'ont été entre 1 et 126 fois sur les 6 années de données (15 fois en moyenne). 64 % à 70 % des 2300 cellules de distribution ont ainsi été inspectées par année

entre 2016 et 2021. Au niveau temporel, 2021 a été l'année avec le moins de campagnes (5 180) et 2017 celle avec le plus de campagnes (5 964), avec en moyenne 5 584 campagnes par an (Figure 20, médiane=5 651). Enfin, les campagnes de recherche de fuite ont duré en moyenne 11 h, avec une forte variabilité (entre 0.25 h et 474 h, soit environ 60 jours de travail).

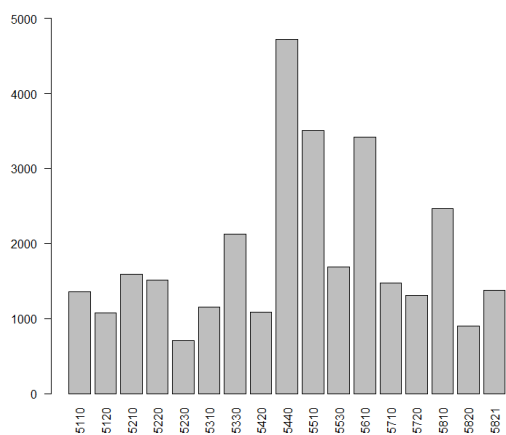


Figure 19 : Nombre de campagnes de RF par secteur de distribution entre 2016 et 2021.

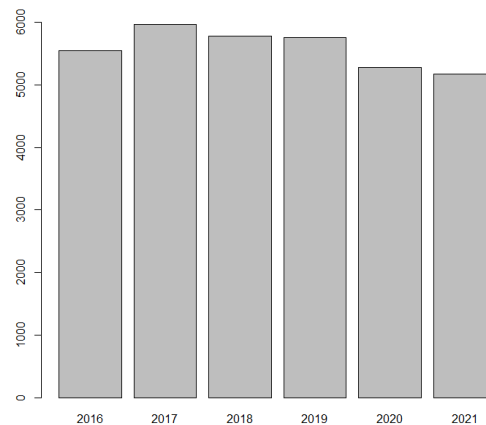


Figure 20 : Nombre de campagnes de RF par année sur toute la SWDE.

8.3.1.2. Succès des campagnes RF

Au total, 98 690 défaillances (fuites + crash, tous types d'activités confondus) ont été détectées et réparées sur tout l'historique des données de la SWDE, entre le 22/03/2001 et le 31/12/2021. Sur notre période d'intérêt de 2016-2021, 58 550 défaillances ont été détectées, dont 23 907 trouvées par la RF (41%). Parmi ces dernières, 83% ont pu être rattachées à des campagnes RF (19 858 défaillances rattachées), c'est-à-dire qu'elles ont vraisemblablement été trouvées grâce à une campagne de RF ; elles représentent 12 508 fuites (63%) et 7 351 crashes (37%).

Remarque : seules 30 339 campagnes RF sont utilisées ici. En effet, 3 163 campagnes existantes via la prestation de SAP se déroulent sur des cellules absentes du SIG en date de 2022, ce qui explique la différence entre le nombre en 8.3.1.1 (33 502) et ici. Ces campagnes sur des cellules « inexistantes » sont donc ignorées par la suite.

Sur les 30 339 campagnes RF considérées sur la période de 2016-2021, 60% ont été totalement « infructueuses », c'est-à-dire qu'elles n'ont trouvé ni fuite, ni crash, et 30% ont trouvé au moins une fuite (Tableau 14).

Tableau 14 : Nombre de campagnes de RF selon leurs résultats.

2016-2021	0 fuite trouvée	Au moins 1 fuite trouvée
0 crash trouvé	18 212 (60%)	7 327 (24%)
Au moins 1 crash trouvé	3 046 (10%)	1 754 (6%)

Cette forte proportion de campagnes infructueuses pose la question d'un effet annuel possible sur le succès des campagnes. En analysant les données par année, 2016 et 2017 montrent une proportion de campagnes infructueuses statistiquement plus élevée que les années 2018 à 2021 (Tableau 15, test de proportion des campagnes infructueuses entre 2016-2017 et 2018-2021 : $\chi^2=88$, $df=1$, $p<2.2 \times 10^{-16}$). Les années 2016 et 2017 ont un taux de succès des campagnes anormalement faible à cause de leur taux plus faible de défaillances déclarées comme trouvées par la RF. Ce dernier est rendu plus faible en 2016 et 2017 par de nombreux OT mal renseignés (absents ou au format incorrect) dans les fichiers fuites, qui empêchent de trouver la défaillance correspondante dans SAP et de vérifier l'attribut « par RF » (en cas d'absence de cette donnée entre 2016 et 2021 inclus, on considère que la défaillance n'est pas trouvée par la RF, voir Annexe 3).

Tableau 15 : Résultat des campagnes de RF. Les lettres indiquent les différences statistiques entre les proportions de campagnes n'ayant rien trouvé (test de proportion deux à deux). Les proportions ayant les mêmes lettres ne sont pas statistiquement différentes entre elles.

	2016	2017	2018	2019	2020	2021
Campagnes n'ayant rien trouvé	3 577 (71%) a	3 555 (65%) b	2 969 (57%) c	2 914 (56%) c	2 590 (54%) c	2 607 (55%) c
Campagnes ayant trouvé quelque chose	1 467	1 884	2 268	2 245	2 164	2 099

Une seconde analyse des données focalisée sur les 19 856 campagnes RF de 2018-2021 indique une baisse de la proportion de campagnes infructueuses à 56% tandis que la proportion de campagnes ayant trouvé au moins une fuite augmente à 34% (Tableau 16).

Tableau 16 : Nombre de campagnes de RF selon leurs résultats.

2018-2021	0 Fuite trouvée	Au moins 1 fuite trouvée
0 crash trouvé	11 080 (56%)	5 498 (28%)
Au moins 1 crash trouvé	2 034 (10%)	1 244 (6%)

De ce fait, les seules années 2018-2021 sont considérées pour la suite, pour l'ensemble du territoire de la SWDE.

Il faut noter que le taux de succès des campagnes est lié à la définition des campagnes (cf seuil de 3 jours ouvrés). Ainsi si le délai interopération était augmenté, le nombre de campagne serait plus faible et le taux de succès des campagnes plus élevé.

8.3.2. Régression logistique

Les données de régression sur toute la SWDE sur 2018-2021 correspondent ainsi à un fichier contenant 3 060 079 lignes, qui correspondent à 213 902 tronçons distincts (rappel : un même tronçon est présent autant de fois qu'il y a eu de campagnes RF sur sa cellule).

Ces tronçons se répartissent ainsi en terme de matériaux :

- Fontes : 53% (113 578 tronçons : 72 229 fontes grises, 41 349 fontes ductiles, 4 centriflex)
- Plastiques : 31% (65 321 tronçons: 55 702 PVC, 9 619 PE)
- AMCI : 10% (20 561)
- Acier : 6% (13 136)
- Autres : < 1% (1 302 tronçons : 1 191 inconnus, 111 en plomb, Béton Âme Tôle, Polyesther ou Inox)

8.3.2.1. Sélection de modèles

Seul l'effet des secteurs de distributions a pu être testé en covariable qualitative fixe (17 modalités), l'effet des cellules (1 865 modalités) ou des tronçons (213 968 modalités) ne pouvant pas être géré par le logiciel (création de n-1 colonnes indicatrices pour n modalités). Pour construire des modèles fixes et mixtes comparables, la seule échelle géographique du secteur de distribution a donc été considérée.

Un modèle complet comprenant les 12 covariables ci-dessous a tout d'abord été testé :

- matériau ;
- diamètre ;
- logarithme naturel du nombre de raccordements + 1 ;
- âge à l'inspection ;
- % de longueur dans classe CLC1 (zone très urbanisée) ;
- % de longueur dans classe CLC5 (zone naturelle) ;
- nombre de défaillances de conduites passées (fuites + crashes de type d'activité SAP 100, 105 et 999) ;
- logarithme naturel de la longueur ;
- interaction matériau : corrosivité ;
- % de longueur sous chaussée ;
- pression ;
- secteur de distribution.

La covariable secteur de distribution a été considérée soit comme une covariable qualitative à 17 modalités dans le modèle fixe, soit comme un facteur aléatoire dans le modèle mixte. Pour les deux types de modèles, les covariables % en CLC1 et % en CLC5 ont été retirées des modèles finaux car non-significatives. De plus, afin de travailler sur des covariables ayant le même ordre de grandeur, ont été utilisés pour les modèles finaux le diamètre en décimètre, l'âge à l'inspection en siècle et la pression en décibars.

8.3.2.2. Comparaison modèle fixe / mixte

Un modèle fixe et un modèle mixte ont ainsi été comparés, comprenant les mêmes 10 covariables (Tableau 17, Figure 21 et Figure 22). La covariable du secteur de distribution a été utilisée comme covariable qualitative dans le modèle fixe, et comme facteur aléatoire sur l'ordonnée à l'origine sur le modèle mixte (le modèle mixte calcule une ordonnée à l'origine –intercept– distincte pour chaque secteur).

Tableau 17 : Comparaison des estimations des paramètres entre modèle fixe et mixte de régression logistique sur la probabilité d'état fuyard. Âge à l'inspection : âge du tronçon au début de l'inspection en siècle. Diamètre en décimètre. Ndef : nombre de défaillances sur conduites (fuites + crashes de type 100, 105 et 999). Pression en décibars. Les valeurs en gras sont les *p*-value significatives, indiquant que la covariable agit significativement sur la probabilité d'état fuyard.

	Modèle fixe	Modèle mixte
--	-------------	--------------

		Estimation	p-value	Estimation	p-value
Intercept		-10.8	$< 2 \times 10^{-16}$	-10.3	$< 2 \times 10^{-16}$
Matériau : Référence = Acier	AMCI	-0.15	0.02	-0.16	0.01
	Autre	0.31	0.25	0.31	0.24
	Fontes	-0.14	5.1×10^{-3}	-0.14	4.4×10^{-3}
	Plastiques	0.29	2.9×10^{-5}	-0.29	2.0×10^{-5}
Age à l'inspection		1.90	$< 2 \times 10^{-16}$	1.89	$< 2 \times 10^{-16}$
Ln(longueur)		0.66	$< 2 \times 10^{-16}$	0.66	$< 2 \times 10^{-16}$
Diamètre		-0.50	$< 2 \times 10^{-16}$	-0.50	$< 2 \times 10^{-16}$
Ndef		0.34	$< 2 \times 10^{-16}$	0.34	$< 2 \times 10^{-16}$
Ln(nb raccordement + 1)		0.14	$< 2 \times 10^{-16}$	0.14	$< 2 \times 10^{-16}$
Pression		0.08	0.01	0.08	0.01
% chaussée		0.19	5.1×10^{-7}	0.19	4.3×10^{-7}
Interaction Matériau : corrosivité	Acier :orros	0.01	0.81	0.02	0.75
	AMCI :orros	-0.04	0.46	-0.04	0.48
	Autres :orros	-0.06	0.76	-0.06	0.77
	Fontes :orros	0.06	3.7×10^{-3}	0.07	2.1×10^{-3}
	Plast :orros	0.08	0.03	0.08	0.02
Secteur de distribution	Effet fixe :	14/16	13/16 secteurs < 0.05		
	Référence = 5110	secteurs > 0			
	Effet aléatoire			Variance 0.25	Ecart type 0.50

Les coefficients sont identiques entre les deux modèles : l'âge à l'inspection, la longueur, le nombre de défaillances passées, le nombre de raccordements, la pression et le pourcentage sous chaussée ont un effet positif sur la probabilité d'état fuyard (coefficients positifs). Cela signifie que plus ces variables augmentent, plus la probabilité de fuir d'un tronçon augmente, donc plus un tronçon est vieux, long, avec beaucoup de défaillances passées et de raccordements, soumis à une forte pression d'eau et ayant un fort pourcentage de sa longueur sous chaussée, plus il est susceptible d'être fuyard. A l'inverse, plus le diamètre du tronçon augmente, plus sa probabilité d'être fuyard diminue. Les tronçons en fonte ou en amiante-ciment fuient moins que ceux en acier, tandis que ceux en plastique

fuients plus, et il n’y a pas de différence significative entre la probabilité d’état fuyard des tronçons en acier et des tronçons dans les autres matériaux (*i.e.* Inconnu, plomb, inox, BAT, polyester).

Le modèle fixe indique que quasiment tous les secteurs de distribution fuient plus que le secteur 5110 (seul le secteur 5230 fuit significativement moins, Figure 21) et le modèle mixte indique que le regroupement des données par secteur a un sens.

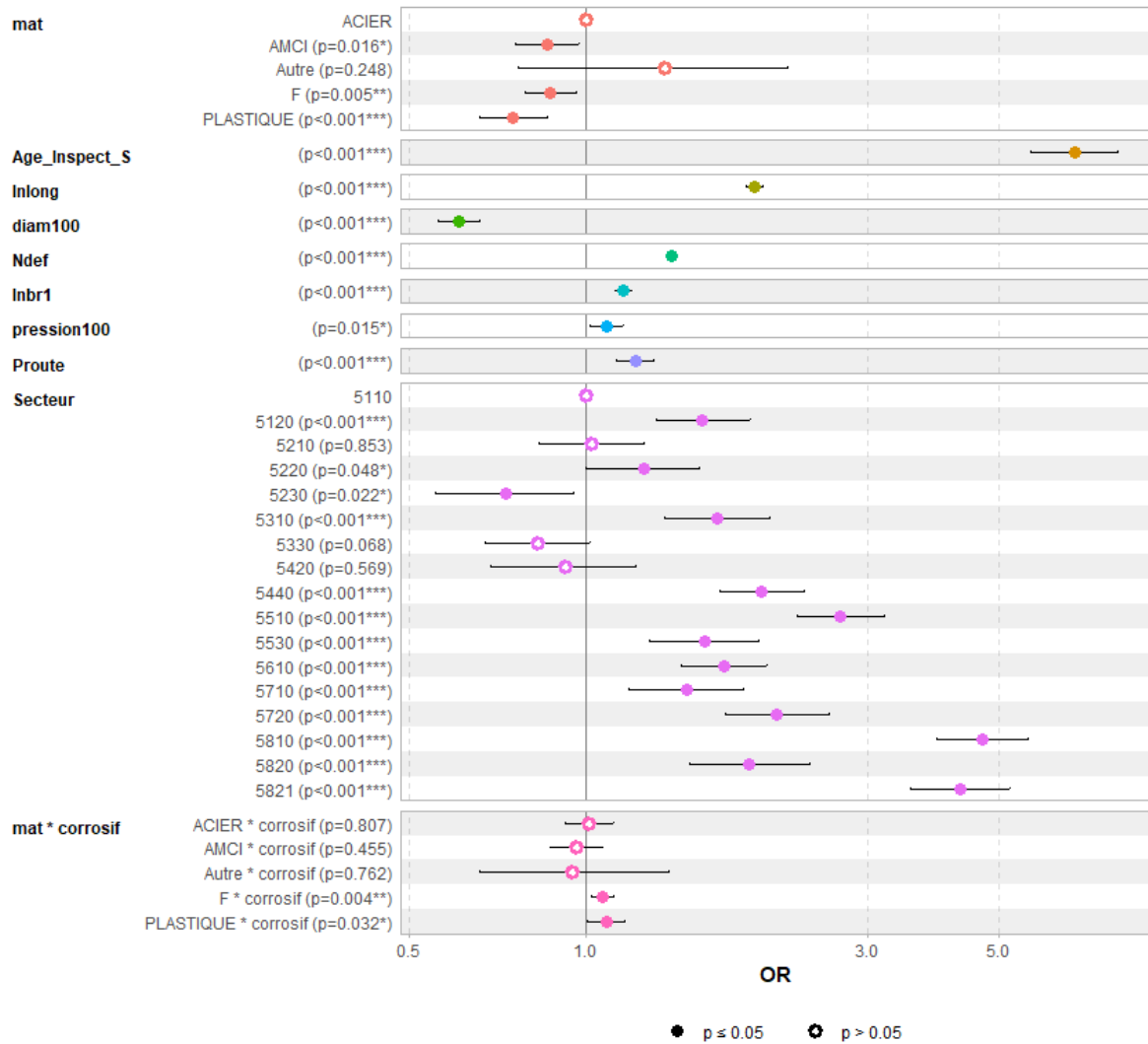


Figure 21 : Odd-ratios des coefficients du modèle fixe et leurs intervalle de confiance à 95%. L’odd-ratio est l’exponentielle de la valeur du coefficient. Mat : Matériau, Age_Inspect_S : âge du tronçon à l’inspection (en siècle). lnlong : logarithme de la longueur du tronçon. Diam100 : diamètre (en décimètre). Ndef : nombre de défaillances passées. lnbr1 : logarithme du nombre de raccords +1. Pression100 : pression (en décibars). Proute : pourcentage de la longueur du tronçon sous la chaussée. Secteur : secteur de distribution. Mat*corrosif : interaction entre matériau et corrosivité de l’eau.

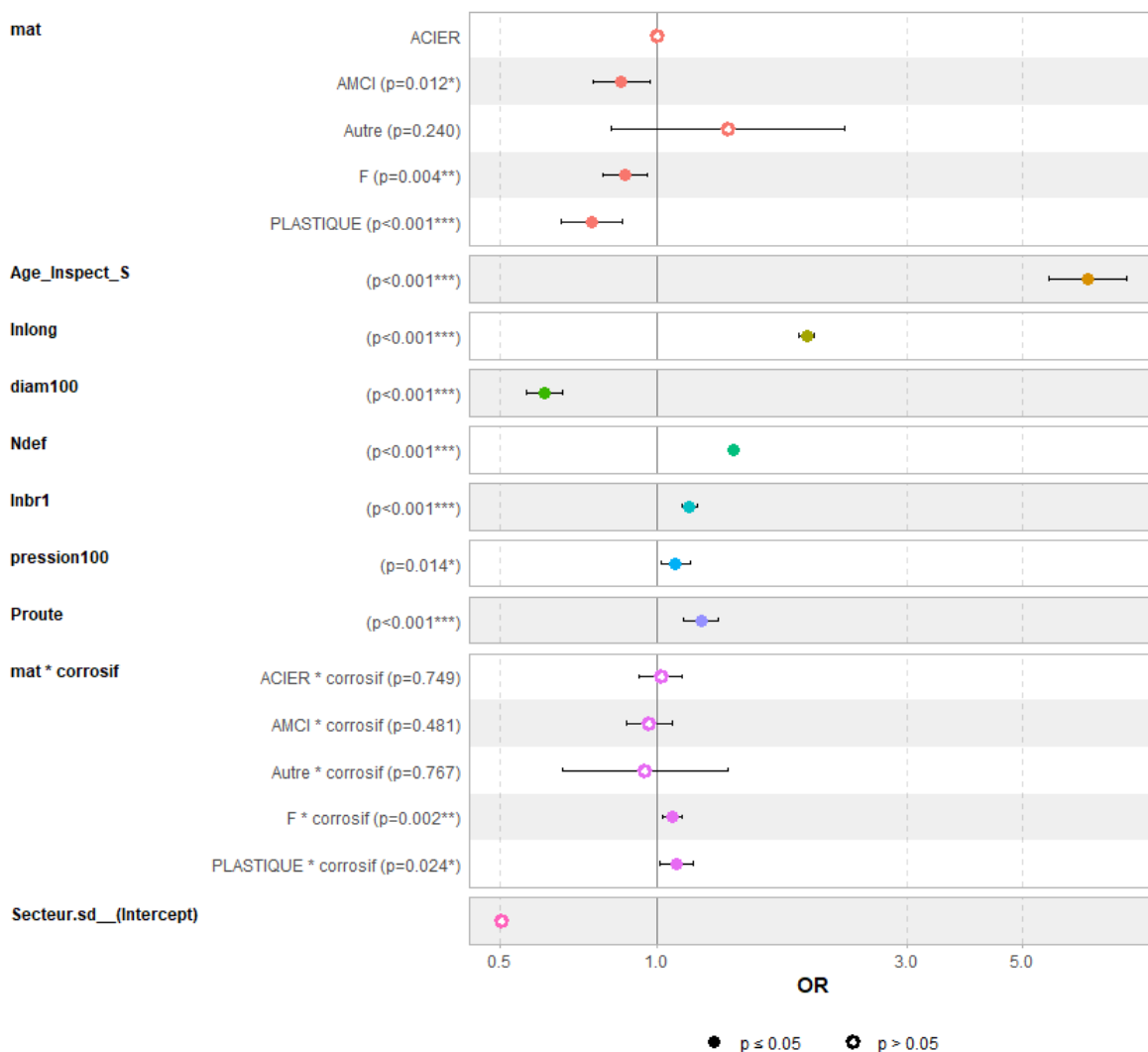


Figure 22 : Odd ratios des coefficients du modèle mixte et leurs intervalle de confiance à 95%. L'odd-ratio est l'exponentielle de la valeur du coefficient. Mat : Matériau, Age_Inspect_S : âge du tronçon à l'inspection (en siècle). lnlong : logarithme de la longueur du tronçon. Diam100 : diamètre (en décimètre). Ndef : nombre de défaillances passées. lnbr1 : logarithme du nombre de raccordements +1. Pression100 : pression (en décibars). Proute : pourcentage de la longueur du tronçon sous la chaussée. Mat*corrosif : interaction entre matériau et corrosivité de l'eau. Secteur.sd_(Intercept) : écart type des différentes ordonnées à l'origine pour chaque secteur de distribution.

Remarque 1 : Il est en théorie plus satisfaisant de considérer le modèle mixte, dans le sens où la non-indépendance des données est une réalité, et où le regroupement des données par secteur de distribution fait sens avec une organisation réelle du travail. Le modèle mixte correspond ainsi mieux à la réalité des données que le modèle fixe de par sa construction même. Cependant, il est plus coûteux en temps de calcul (1h15 contre 2min sur la même machine pour les modèles présentés), et l'équivalence des résultats des deux modèles (mixtes et fixes) montre que le modèle fixe, plus praticable à l'usage, n'induit en outre pas de biais d'estimation des effets.

Remarque 2 : L'impact des covariables sur la probabilité d'état fuyard est identique entre modèle LEYP(2s) et modèle de régression logistique. En effet, quand les variables sont significatives, elles impactent l'intensité de défaillances (pour le LEYP) et la probabilité d'état fuyard (pour la régression logistique) dans le même sens. Par exemple, plus une conduite est longue, plus elle est susceptible

d'accumuler des défaillances et d'être en train de fuir. Si cela paraît intuitivement logique, il est intéressant de l'avoir montré et cela pourra potentiellement simplifier le travail de réunion des deux modèles dans le futur (Tâche 09).

8.3.2.3. Probabilité d'état fuyard d'un tuyau sachant qu'il a été inspecté

Grâce au modèle mixte, la probabilité pour chaque tronçon d'être fuyard au 1^{er} janvier 2022 a été calculée, soit juste après la fin des observations. Il apparaît qu'un nombre très restreint de tronçons ont une probabilité d'être fuyard supérieure à 5 % (123 tronçons), tandis que la très grande majorité ont une probabilité très faible (Figure 23 et Figure 24 à Figure 27).

Il est important de noter ici une différence majeure entre modèles LEYP et régression logistique : si le modèle LEYP est capable par construction de créer des prédictions sur un intervalle de temps donné (comme 2022-2026 présenté en paragraphe 4.3.1.3 et en Annexe 7), c'est impossible pour le modèle de régression logistique. Par nature, les régressions ne peuvent produire une prédiction qu'à un temps t précis, à condition d'être capable de prédire la valeur des covariables pour ce temps t spécifique. Le phénomène à prédire est en outre conditionnel à l'inspection à l'âge t . Cela pose ainsi un problème théorique complexe quant à la prédiction du nombre de défaillances utilisé dans les modèles présentés. La piste du calcul de ce nombre de défaillances théoriques à t pour chaque tronçon via un modèle LEYP est actuellement explorée, mais la formalisation du couplage entre les modèles LEYP et ceux de régression est complexe et fait l'objet du travail de la Tâche 09.

Ainsi, les prédictions présentées ici sont des prédictions au 1^{er} janvier 2022, soit le lendemain de la fin des observations, où il peut être raisonnablement supposé que le nombre de défaillances des tronçons n'a pas bougé significativement depuis la veille.

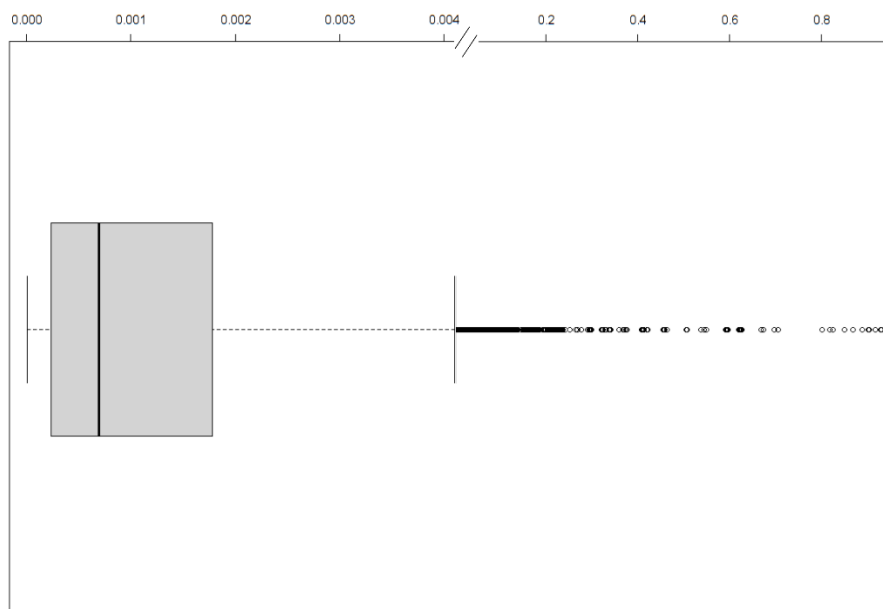


Figure 23 : Distribution de la probabilité d'état fuyard des tronçons au 1^{er} janvier 2022 issue du modèle mixte. Merci de noter la rupture d'échelle au niveau de la valeur 0.004 (symbole //) pour faciliter la représentation.

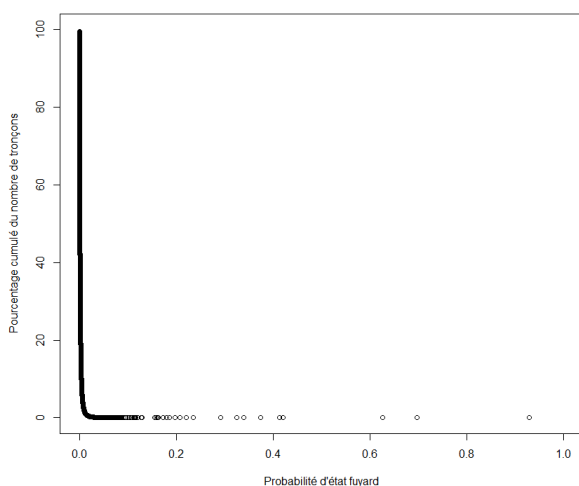


Figure 24 : Probabilité d'état fuyard selon le pourcentage d'effectif cumulé de tronçons.

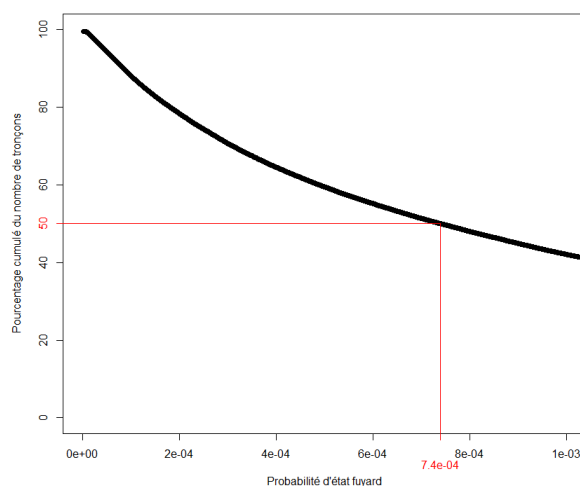


Figure 25 : Probabilité d'état fuyard selon le pourcentage d'effectif cumulé de tronçons. Zoom. La probabilité en rouge correspond à la probabilité de 50% de l'effectif cumulé.

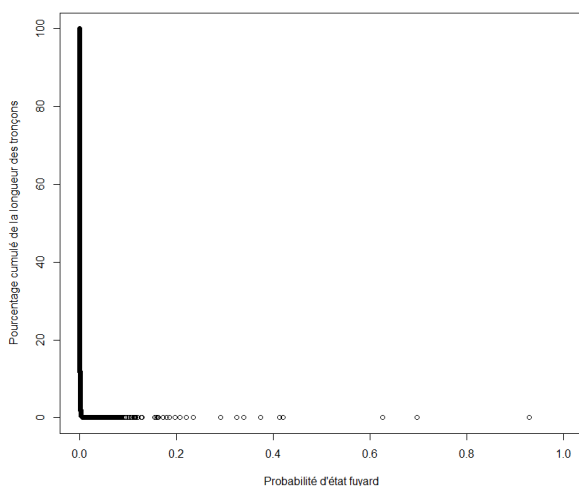


Figure 26 : Probabilité d'état fuyard selon le pourcentage de longueur cumulée des tronçons.

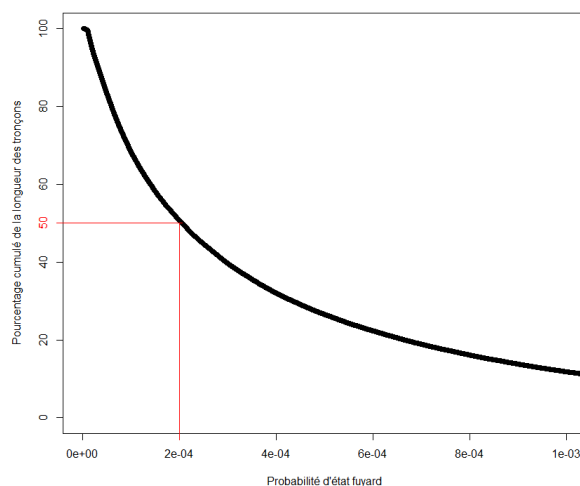


Figure 27 : Probabilité d'état fuyard selon le pourcentage de longueur cumulée des tronçons. Zoom. La probabilité en rouge correspond à la probabilité de 50% de la longueur cumulée.

Cette distribution de la probabilité d'état fuyard est très asymétrique (quelques centaines de tronçons avec des probabilités très élevées et le reste avec des probabilités quasi nulles). Pour mieux en comprendre les raisons, un examen minutieux de la centaine de tronçons avec des probabilités très élevées reste à réaliser. Les valeurs des paramètres montrent en première approche que l'effet de l'âge du tronçon, combiné à sa longueur et à son historique de défaillances sont des déterminants importants des résultats. Cela interroge ainsi sur la manière de prendre en compte (ou non) un historique de défaillance dans ce modèle, et lequel. Les modèles présentés se sont limités à la somme du nombre de défaillances des conduites au cours du temps, mais d'autres possibilités sont envisageables comme un historique prenant en compte toutes les défaillances de tous types d'activités

confondus, ou bien uniquement les fuites. Ceci fera l'objet d'explorations futures dans le cadre de cette tâche.

8.4. Objectifs futurs

8.4.1. Amélioration de l'hypothèse d'inspection des tronçons

Actuellement, afin de pouvoir faire tourner les modèles de régressions logistiques, l'hypothèse est faite qu'une campagne de recherche de fuite inspecte tous les tronçons d'une cellule. Or, (Paragraphe 8.3.1.1) une campagne dure en moyenne 11 h. Si on estime qu'un agent (ou un binôme d'agent) inspecte en moyenne 40 km/semaine, ce qui revient à environ 1 km/h, cette hypothèse d'inspection totale de la cellule durant la campagne ne tient plus pour les cellules de plus de 11 km (38 % des cellules). Il serait donc intéressant de voir quels éclairages peuvent nous être donnés par la Tâche 06 « Analyse des pratiques de recherche de fuite », pour revisiter cette hypothèse, voire la conforter en montrant que ne sont pas inspectées en pratique les portions de la cellule où les chances de détecter une fuite sont quasi-nulles.

8.4.2. Exploration de la prédiction de la probabilité d'état fuyard

Comme énoncé dans le paragraphe 8.3.2, deux grandes pistes de travail sont actuellement en train d'être explorées :

1. L'impact de l'historique des défaillances sur le calcul de la probabilité d'état fuyard (voir la conclusion du paragraphe 8.3.2.3)
2. La prédiction de cet historique dans le futur, qui nécessite un couplage formel entre modèle de régression logistique et LEYP, qui est l'objet entier de la Tâche 09 (paragraphe 8.3.2.3) .

8.4.3. Probabilité d'état fuyard non conditionnée par l'inspection

Afin de déterminer la probabilité qu'un tronçon soit dans un état fuyard en s'affranchissant de son inspection, c'est-à-dire le calcul de $P(Y = 1|Z, X)$, les quatre calculs de probabilités suivantes sont nécessaires (formule des probabilités totales) :

$$P(Y = 1|Z, X) = P(Y = 1|D = 1, Z, X) \times P(D = 1|X) + P(Y = 1|D = 0, Z, X) \times P(D = 0|X)$$

La modélisation actuelle permet d'évaluer $P(Y = 1|D = 1, Z, X)$, et sont donc encore à déterminer $P(Y = 1|D = 0, Z, X)$ et $P(D = 1|X)$ (sachant que $P(D = 0|X) = 1 - P(D = 1|X)$).

$P(D = 1|X)$ correspond à la probabilité qu'une cellule soit inspectée par la RF selon ses caractéristiques X . Cette probabilité devrait pouvoir être estimée grâce aux résultats de la Tâche 05 sur la sectorisation.

$P(Y = 1|D = 0, Z, X)$ correspond à la probabilité qu'un tronçon soit dans un état fuyard **sachant** que la cellule n'a **pas** été inspectée, autrement dit sachant que la cellule ne présente pas de caractéristiques nécessitant l'envoi d'une équipe de recherche de fuite (par exemple pas de dérive de débit de nuit importante). L'évaluation de cette probabilité nécessiterait ainsi des expérimentations où des équipes de recherche de fuite seraient envoyées sur des cellules ne présentant pas de signes de dérive de débit.

9. Tâche 08 – Analyse de la vulnérabilité des usagers

Cette section a pour objectif de présenter la méthodologie mise en place pour capturer la vulnérabilité des consommateurs au risque de sécheresse dans la Tâche 08. Elle ne proposera pas à ce stade du projet de résultats, le questionnaire n'ayant été clôturé que le 1^{er} septembre 2022 dans la matinée.

9.1. Contexte et problématique

En Wallonie, le déficit de précipitations des années récentes (et de façon particulièrement intense en 2018 et 2020) occasionne des périodes de sécheresse, avec des conséquences observables sur les ressources tels que des étiages parfois sévères et une baisse des niveaux de certains barrages et des nappes souterraines. Les épisodes de sécheresse ont ainsi de multiples impacts sur les écosystèmes (dégradation de la qualité des cours d'eau et de concentration en nitrate des eaux souterraines). Ils s'accompagnent également d'une hausse de la demande en eau de tous les acteurs, dont les ménages. Des pénuries d'eau potable ont ainsi déjà été observées localement et certaines communes mettent régulièrement en place des arrêtés de police visant à limiter l'usage de l'eau du robinet (par exemple interdiction de lavage des véhicules, de remplissage des piscines et citernes, de l'arrosage des jardins d'agrément). Or la transposition au niveau belge des scénarios du GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat) fait craindre une augmentation de la fréquence de ces épisodes de sécheresse en été.

Dans ce contexte, le premier objectif de cette tâche est d'apprécier la vulnérabilité des consommateurs au risque de sécheresse, mais aussi d'estimer leur consentement à payer pour des mesures visant à en limiter les conséquences sur l'approvisionnement en eau potable. Ce travail se concentre exclusivement sur les particuliers. Il repose sur une enquête auprès de ces derniers. Il s'agit aussi de comparer la vulnérabilité « objective » telle que définie par la SWDE et la vulnérabilité « subjective » telle que ressentie par les ménages. À cet effet, trois groupes de communes wallonnes ont été définis (Tableau 18) :

1. Le premier regroupe des communes dans lesquelles les problèmes de sécheresse ont des conséquences sur la distribution qui se sont traduites au moins une fois par des mesures de restriction des usages depuis janvier 2019 ;
2. Le second regroupe des communes dans lesquelles la sécheresse a des conséquences sur le fonctionnement de la SWDE (il est par exemple nécessaire d'approvisionner par camions des châteaux d'eau afin de garantir la continuité du service) mais de ce fait sans conséquences importantes pour les clients ;
3. Dans le troisième groupe, la sécheresse n'a aucune conséquence ou est inexistante (présence de barrages, de captages importants ou de liaisons par exemple). Ces communes ont été choisies à proximité de celle du groupe 1 afin que les contextes pluviométriques soient néanmoins relativement « proches ».

Les communes de chaque zone ont été sélectionnées après discussion avec les services de la SDWE concernés. Nous avons imposé pour contrainte que la totalité du territoire de chacune soit situé dans la même Zone de Distribution en Eau (ZDE) afin de nous assurer que la qualité de l'eau puisse être considérée comme uniforme et surtout provienne de la même source. Dans certaines communes, l'ensemble des habitants ne sont pas desservis par la SWDE mais ceci n'est pas problématique en raison du protocole de collecte des données que nous avons adopté (cf. plus bas)

Tableau 18 : Liste des communes d'enquête.

	Groupe 1	Groupe 2	Groupe 3
Communes	Beauraing* Ciney* Durbuy* Gesves Ohey Rochefort* Vielsalm	Beloeil* Bertrix Binche Braine-l'Alleud* Braine-le-Comte Braives Crisnée Daverdisse Gerpennes Grez-Doiceau Hannut Jalhay La Louvière Mons Paliseul Plombières Rebecq Remicourt Saint-Vith* Sprimont Waimes* Waremme Yvoir	Andenne Assesse Ath Aywaille Dinant Eupen Ferrières Herve Houyet Lessines Leuze-en-Hainaut Lierneux Marche-en-Famenne Namur Raeren Soumagne Thimister-Clermont Tournai Wellin

Note : * communes non entièrement desservies par la SWDE

9.2. Construction du questionnaire

Le questionnaire d'enquête a été construit autour de questions d'évaluation monétaire, visant à estimer le consentement à payer (CAP) des usagers. En raison du caractère multidimensionnel des conséquences des sécheresses, nous avons fait le choix de mobiliser ici la Méthode des Choix Multi-Attributs (MCMA) (*Choice Experiment* en anglais). Cette approche permet en effet de décomposer le bien ou service sur lequel porte l'étude (ici l'approvisionnement en eau potable), en un certain nombre de caractéristiques. L'analyse économétrique des données collectées lors de l'enquête fournit alors à la fois les CAP pour chaque caractéristique et les arbitrages entre ces dernières.

Suite à une revue de la littérature sur l'évaluation des CAP pour lutter contre les effets des sécheresses, deux scénarios distincts ont été construits, qui ont été présentés de manière aléatoire aux répondants : le premier décline les conséquences des sécheresses en termes de restrictions des usages et le second en termes de performance de service. Dans les deux cas, il est supposé que différentes mesures de prévention peuvent être mises en œuvre pour lutter contre les conséquences des sécheresses (remplissage des châteaux d'eau par des camions citernes ; pose de conduites d'interconnexion permettant la jonction entre les zones disposant d'importantes ressources et celles aux ressources plus limitées ; recherche de nouvelles sources d'eau potable (nouveaux captages, dessalement d'eau de mer...) ; développement de projets de ressources alternatives (réutilisation des eaux de pluie ou des eaux usées traitées...) ; etc.). Ces mesures, mises en œuvre et coordonnées par la SWDE, la Région wallonne et les communes, s'accompagneraient d'une communication accrue auprès de l'ensemble des acteurs sur la nécessité de faire attention à sa consommation d'eau potable.

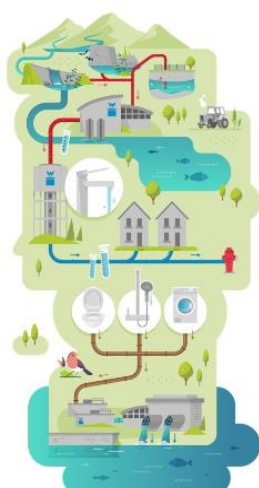
9.2.1. Le scénario « Restriction des usages »

Le premier scénario suppose que selon les mesures mises en œuvre et leur ampleur, les restrictions des usages de l'eau du robinet sont susceptibles de varier dans les 10 prochaines années en termes de :

- Durée des restrictions pour les ménages, qui pourrait augmenter de 10, de 15, de 20 voire de 25 jours par an par rapport à aujourd'hui ;
- Nombre de communes wallonnes touchées, qui pourrait passer d'1 sur 20 actuellement à 1 sur 10, voire 1 sur 4 ou même 1 sur 2 ;
- Type de restrictions d'usage imposées aux ménages. Cela pourrait commencer par l'interdiction du nettoyage des véhicules ou façades et terrasses. Puis s'ajouteraient une interdiction du remplissage des bassins, piscines et mares ou des citernes et enfin de l'arrosage des pelouses, jardins et potagers ou des jeux d'eau d'extérieur ;
- Entretien des espaces verts situés sur les espaces publics (pelouses, terrains de sport, plantes situées sur les ronds-points par exemple). Il s'agirait soit de maintenir leur qualité esthétique et paysagère, soit de raisonner l'arrosage de façon à garantir leur survie mais avec des conséquences esthétiques temporaires, soit de ne plus les arroser du tout en période de sécheresse.

Ces choix sont supposés engendrer des augmentations du prix de l'eau qui sont variables. La cinquième caractéristique est donc monétaire. Il a été précisé dans le questionnaire que les fonds collectés seraient exclusivement utilisés pour ce projet.

Les enquêtés se voient alors proposer six cartes de choix combinant les différentes caractéristiques prises à différents niveaux, i.e. décrivant des scénarios possibles de restrictions des usages de l'eau du robinet dans leur commune en période de sécheresse à l'horizon 2030 (voir la Figure 28 pour un exemple)¹. Tout ce qui n'est pas mentionné dans les options reste identique à la situation actuelle (pression et qualité sanitaire de l'eau distribuée par exemple). Ces cartes de choix ont été construites grâce au logiciel dédié NGENE. Dans chacune, le répondant doit choisir dans chaque carte l'option qu'il préfère. L'option C reste inchangée d'une carte à l'autre et correspond à un maintien de la gestion actuelle des sécheresses. Elle est proposée à un coût nul (0 €). Il est supposé que les mesures proposées s'imposeraient à l'ensemble des habitants de la commune afin d'éviter les comportements stratégiques.


















	Option A	Option B	Option C (sans mesure supplémentaire)
Durée des restrictions	25 jours de plus 	10 jours de plus 	25 jours de plus 
Nombre de communes touchées	1 commune sur 20 	1 commune sur 4 	1 commune sur 2 
Type de restrictions	Ni nettoyage, ni remplissage, ni arrosage 	Pas de nettoyage 	Ni nettoyage, ni remplissage, ni arrosage 
Entretien des espaces verts	Arrosage classique 	Arrosage raisonné 	Pas d'arrosage 
Augmentation de la facture d'eau	36 € par an 	48 € par an 	0 € par an 

Figure 28 : exemple de carte de choix du scénario « Restrictions des usages »

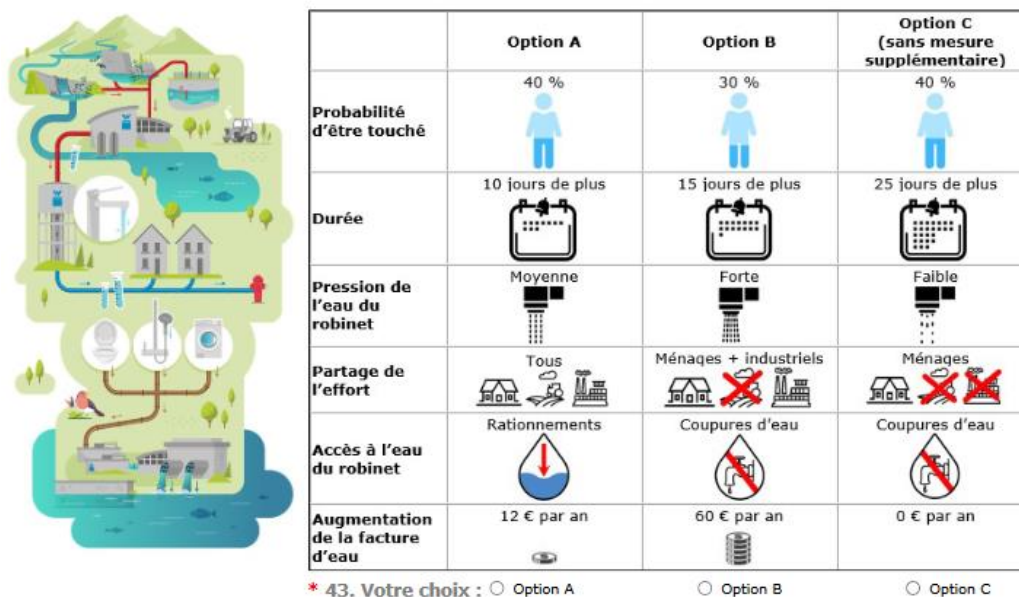
¹ Il est à noter que pour des raisons statistiques, 12 séries de choix ont été générées pour chaque scénario qui ont ensuite été regroupées en deux blocs de six, répartis de manière aléatoire entre les répondants.



















9.2.2. Le scénario « Performance du service rendu »

Selon le même principe, le second scénario considère que selon les mesures mises en œuvre et leur ampleur, la politique de prévention des conséquences des sécheresses est susceptible dans les 10 prochaines années de :

- Réduire la durée des conséquences des épisodes de sécheresse qui pourrait augmenter de 10, de 15, de 20 voire de 25 jours par an par rapport à aujourd'hui ;
- Maintenir la pression de l'eau du robinet qui pourrait être moyenne, avec des conséquences en termes de confort (filet d'eau sous la douche...) et d'efficacité (allongement du temps d'arrosage...), voire basse (pas d'eau à l'étage, impossibilité d'utiliser les machines à laver le linge ou la vaisselle...) ;
- Moduler les acteurs visés par le partage de l'effort de réduction des consommations d'eau du robinet. Il pourrait s'agir soit des seuls ménages, soit des ménages et agriculteurs, soit des ménages et industriels, soit encore de tous ces acteurs ;
- Maintenir un accès à l'eau du robinet pour les ménages. Outre les restrictions, il est possible que les quantités d'eau à disposition soient rationnées, voire que les ménages aient à faire face à des coupures d'eau ponctuelles (par ex. 1 jour par semaine sur les mois de juillet et août) ;
- Réduire la probabilité d'être touché(e) par ces conséquences des sécheresses, qui pourra être de 10 %, de 20 %, de 30 % ou de 40 % (soit en moyenne 1, 2, 3 ou 4 années sur 10).

Ici aussi, les choix engendreraient des augmentations du prix de l'eau qui sont variables. Les hypothèses et la tâche des enquêtés sont les mêmes que ceux décrits précédemment pour le scénario « Restrictions des usages ». Un exemple de série de choix est proposé dans la Figure 29.



	Option A	Option B	Option C (sans mesure supplémentaire)
Probabilité d'être touché	40 % 	30 % 	40 % 
Durée	10 jours de plus 	15 jours de plus 	25 jours de plus 
Pression de l'eau du robinet	Moyenne 	Forte 	Faible 
Partage de l'effort	Tous 	Ménages + Industriels 	Ménages 
Accès à l'eau du robinet	Rationnements 	Coupures d'eau 	Coupures d'eau 
Augmentation de la facture d'eau	12 € par an 	60 € par an 	0 € par an 

* 43. Votre choix : Option A Option B Option C

Figure 29 : exemple de carte de choix du scénario « Performance du service fourni »

9.2.3. Structure du questionnaire

Au final, le questionnaire est construit autour de six blocs de questions traitant :

1. Des aspects socio-démo-géographiques (commune de résidence, genre, âge, type de logement, catégories socio-professionnelles, revenus, etc.) ;
2. De la ressource en eau en général (enjeux autour de la gestion de l'eau, menaces sur la ressource selon le répondant, etc.) ;
3. Des usages en eau, avec notamment des questions sur les pratiques d'économies d'eau ;

4. Du service d'eau du robinet (origine de l'eau distribuée, organisme en charge de cette distribution, critères importants en termes de performance du service, appréciation de cette performance, etc.) ;
5. Du risque de sécheresse, afin d'approcher la perception de ce risque par les ménages mais aussi leur sensibilisation (au sens anglais de *awareness*). Ce bloc contient notamment des questions sur l'information dont ils disposent, leur manière de se la procurer, etc. ;
6. Des séries de choix et questions de suivi associées qui ont pour objectif de s'assurer de la pertinence et de la crédibilité des scénarios proposés.

Chaque scénario est déroulé de la même façon à savoir qu'il commence par une présentation générale des conséquences des sécheresses puis des mesures mises en œuvre par la SWDE pour éviter ses conséquences sur l'approvisionnement en eau potable. Sont ensuite décrites les différentes caractéristiques utilisées dans les cartes de choix. Enfin, la contrainte budgétaire est rappelée, ainsi que cela est préconisé dans la littérature, de même que le montant moyen annuel de la part « eau potable » d'un ménage wallon de trois personnes afin que le répondant dispose d'une référence².

Le questionnaire est proposé en annexe AX5 séparée du présent rapport.

9.3. Protocole d'enquête et collecte des données

La SWDE disposant des adresses mails d'environ la moitié de ses clients, équitablement répartis sur son territoire, nous avons fait le choix d'une passation par internet : 15 000 ménages habitant dans chaque groupe de communes (donc 45 000 au total) ont été sélectionnés de manière aléatoire dans la base de données « clients » et ont reçu de la part de la SWDE un message les invitant à répondre à l'enquête (Figure 30). Selon leur rang (date-heure) de connexion, la version « Restrictions des usages » ou la version « Performance du service rendu » leur ont alors été attribuées de manière aléatoire.

² La littérature montre en effet que la grande majorité des ménages ne connaît pas le montant de sa facture d'eau.



Figure 30 : mail envoyé par la SWDE à un échantillon aléatoire de clients habitant dans les communes d'enquête

Le questionnaire a donc été auto-administré. Il a été testé auprès de 3 000 ménages (1 000 par groupe de communes) en condition d'enquête en juin 2022 puis ajusté à la marge. L'envoi définitif a été réalisé le 20 juillet 2022, avec une relance le 1^{er} août et une autre le 16 août. La date limite de réponse était fixée au 31 août 2022.

Au final, 45,4 % des mails ont été ouverts lors de l'envoi initial du 20 juillet, 50,7 % lors de la première relance et 46,9 % lors de la seconde. Sur les personnes qui l'ont ouvert, un peu plus de 4 360 ont cliqué sur le lien proposé vers le questionnaire hébergé sur le site du logiciel dédié Sphinx, dont 73 % ont complété au moins une question. Le taux de retour est d'environ 15 % si on considère uniquement les mails ouverts. Certaines personnes ayant ensuite arrêté le questionnaire en cours, nous disposerons d'un peu plus de 1 000 réponses pour chacun des scénarios (soit 6 000 choix) exploitables pour l'analyse économétrique, analyse qui est en cours.

10. Tâche 13 – Coordination du Projet et valorisation

10.1. Animation

L'instance centrale d'animation du projet est le comité de suivi opérationnel (CSO) qui est composé des principales personnes de SWDE et INRAE qui sont en charge de l'exécution du projet. Il a tenu six réunions au cours de la seconde année du projet, toutes en visioconférence à l'exception de la dernière, le 1^{er} juin 2022, qui s'est tenue dans les locaux de SWDE à Loyers (proche de Namur).

A deux reprises (en décembre 2021 et en juin 2022) les réunions du CSO ont été complétées par des réunions thématiques associant un nombre de personnes restreint pour approfondir des questions plus précises. Par ailleurs, un sous-groupe spécifique a été institué début 2022 pour traiter plus particulièrement des questions attachées à l'intégration des outils du projet dans le système d'information de la SWDE (tâche 4).

Le comité de pilotage stratégique (CPS), en charge de veiller au bon déroulement du projet et composé de représentants des directions des deux partenaires, s'est réuni en visioconférence en décembre 2021.

10.2. Moyens humains affectés par INRAE au projet

Au cours de la deuxième année du projet, outre les agents permanents mentionnés comme référents dans le Tableau 1, INRAE mobilise trois agents non-permanents et trois stagiaires réalisant leur stage de fin d'étude en Master II ou école d'ingénieur (Tableau 19).

Tableau 19 Personnels non-permanents et stagiaires INRAE

Profil	Mois	Personne recrutée	Début	Fin
Ingénieur de recherche en eau potable	24	Nicolas Rodriguez	01/12/2020	30/11/2022
Post doctorant en sociologie	12	Selma Baati	01/09/2021	31/08/2022
Post doctorant en statistiques	24	Marie Collet	01/09/2021	31/08/2023
Stagiaire Ingénieur eau potable	6	Aurélien Mirebeau	01/03/2022	31/08/2022
Stagiaire Master en sociologie	6	Simon Fidelle	01/04/2022	30/09/2022
Stagiaire Master en économie	6	Achille Payant	01/04/2022	30/09/2022

Pour la suite du projet il est à noter que :

- du fait de son départ en retraite, la contribution d'Yves Le Gat prendra fin au 31/12/2022 ;
- la participation de Marie Collet sera suspendue de septembre 2022 jusqu'à mi-janvier 2023 en raison d'un congé maternité ;
- le contrat de Selma Baati a été prolongé de six mois (nouvelle échéance le 28/02/2023), pour mieux valoriser le matériel recueilli sur le terrain dans le cadre de la tâche 6 ;
- le recrutement d'un ingénieur d'étude en économie pour 6 mois à compter du 01/10/2022 a été lancé pour contribuer à la tâche 8 ;
- le contrat de Nicolas Rodriguez va être prolongé de 9 mois pour maintenir sa contribution au projet jusqu'à son échéance le 31/08/2023.

10.3. Valorisations

Le projet a fait l'objet de deux présentations dans des conférences en 2022.

A partir des résultats des travaux réalisés dans le cadre de la tâche 1, Nicolas Rodriguez a présenté une communication intitulée « Méthodes automatisées de préparation des données pour la gestion patrimoniale du réseau de la SWDE » au 101^{ème} congrès de l'ASTEE qui s'est tenu à Dunkerque du 14 au 16 juin 2022. La présentation (Rodriguez et al, 2022) est jointe au présent rapport (annexe séparée AX6). Dans la continuité de cette intervention, un article des mêmes auteurs et portant le titre « Méthodes automatisées de préparation des données pour la gestion patrimoniale des réseaux d'eau potable » a été soumis pour publication à la revue TSM (Techniques Sciences Méthodes), il est en cours de révision à la date d'établissement du présent rapport.

Lors de la conférence internationale LESAM (Leading-Edge Strategic Asset Management) de l'IWA qui s'est tenue à Bordeaux du 11 au 13 mai 2022, Marie Collet a présenté une communication intitulée « Understanding the leakage process for multi-scale water infrastructure asset management: necessity for a dialogue between sociological and data sciences ». Celle-ci retrace l'articulation entre les travaux des tâches 1, 6 et 7. Le résumé associé à cette intervention (Collet et al., 2022) est proposé en annexes séparées AX7 et AX8 du présent rapport.

10.4. Planning prévisionnel ajusté

Le planning prévisionnel du projet est adapté en fonction de l'avancement réel des travaux et des contraintes connues ou prévisibles qui peuvent l'influencer. La Figure 31 présente le planning prévisionnel dans sa version actualisée à la date du 29/07/2022.

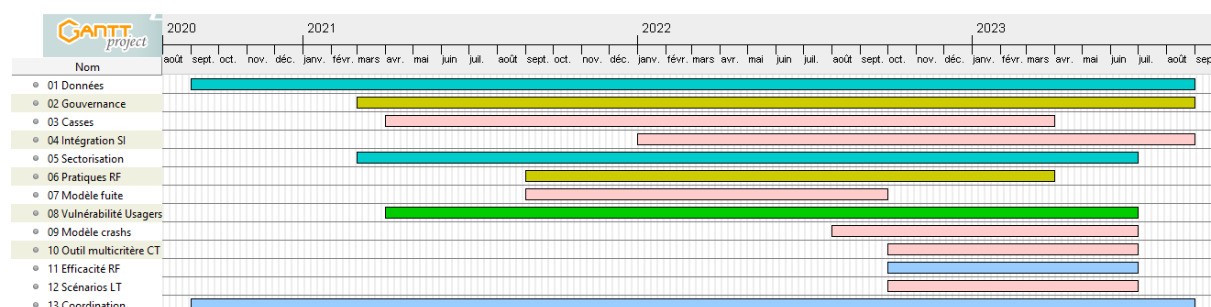


Figure 31 Planning prévisionnel ajusté au 29/07/2022

11. Conclusion

De avancées notables sur de nombreuses tâches du projet GePaME ont été réalisées entre août 2021 et septembre 2022. Le recrutement de 2 CDDs en automne 2021 et de 3 stagiaires au printemps 2022 au sein d'INRAE a constitué un renfort important des moyens humains employés pour l'avancement du projet. On note en particulier les progressions suivantes :

- En tâche 01, la méthode d'affectation des défaillances historiques aux tronçons du SIG a été validée et son principe général a été résumé dans une publication (en cours de révision). L'application des méthodes de traitement des données à l'échelle de la SWDE entière a été réussie. Le traitement des données de RF a permis à la tâche 07 de débuter.
- En tâche 03, les fichiers des cassettes et des tronçons produits en tâche 01 ont rendu possibles les simulations avec le modèle LEYP suivies de la production de cartes de prédiction des défaillances sur Mons, Ath, et Gaurain, incluant les incertitudes sur la localisation des défaillances historiques. Ces incertitudes ont vocation à être comparées avec l'expérience des agents de terrain de la SWDE.

- Le travail en tâche 05 a permis de tracer l'évolution historique du périmètre des 2 300 cellules de distribution et de calculer les débits de ces dernières sur les périodes de validité les plus longues possibles. Le calcul de covariables additionnelles pour les tronçons et leur agrégation à l'échelle des cellules a été essentiel pour les avancées des tâches 03 et 07 (modèles LEYP et régression logistique). La typologie des cellules est en cours de construction, avec des premiers indicateurs patrimoniaux, d'exploitation, et d'environnement déjà calculés à l'échelle des cellules.
- En tâche 07, deux types de modèles de régression logistique ont pu être rapidement développés sur les données traitées de RF de la SWDE. Le travail à venir va permettre de revenir éventuellement sur certains choix de conceptualisation nécessaires pour cette modélisation.
- De nombreuses données ont été collectées et sont en cours de traitement concernant les tâches 02 (entretiens avec les acteurs impliqués dans la gouvernance territoriale de la GPI de la SWDE), 06 (analyse de documents, observations *in situ*, et entretiens des acteurs impliqués dans la RF), et 08 (construction du questionnaire avec la SWDE, envoi et collecte des réponses). Les interactions avec les autres tâches du projet ont amélioré notre compréhension des processus opérationnels et de décision au sein de la SWDE, ce qui a notamment permis une meilleure appropriation des données techniques.
- Le transfert des méthodes de traitement des données et de leur exploitation (dont modèle LEYP) a débuté (tâche 04), grâce à de nombreux échanges et à la mise en place d'un groupe de travail spécifique INRAE-SWDE.
- Le couplage entre tâches se poursuit dans l'optique de progresser en tâche 09 et de démarrer prochainement les tâches 10 et 11.

On note également que plusieurs opportunités de valorisation scientifique des résultats du projet ont été saisies, avec deux participations à des congrès abordant le thème de la GPI (ASTEE et LESAM) et un article soumis à un journal scientifique et technique (TSM). Ces opportunités vont se multiplier au cours de la dernière année du projet (on note déjà une participation prévue au congrès ASTEE de 2023).

12. Références

Aquawal, 2020. Statistiques de l'eau potable et de l'assainissement des eaux usées en Wallonie – Rapport 2020.

Aubin, D., Varone, F., 2001. La gestion de l'eau en Belgique. Analyse historique des régimes institutionnels (1804-2001), Courrier Hebdomadaire du CRISP, n°1731-1732, p. 5-75.

Belghaddar Y., Chahinian N., Seriai A., Begdouri A., Abdou R., Delenne C., 2021. Graph Convolutional Networks: Application to Database Completion of Wastewater Networks. *Water*. 13(12):1681.

Betrie, G.D., Sadiq, R., Tesfamariam, S., et al., 2016. On the Issue of Incomplete and Missing Water-Quality Data in Mine Site Databases: Comparing Three Imputation Methods. *Mine Water Environ* 35, 3–9.

Collet, M., Rodriguez, N., Baati, S., Husson, A., Renaud, E., Caillaud, K., Le Gat, Y., 2022. UNDERSTANDING THE LEAKAGE PROCESS: A DIALOGUE BETWEEN SOCIOLOGY AND DATA SCIENCES. Leading Edge conference for Strategic Asset Management (LESAM), Bordeaux, 11-13 mai 2022.

Denis, J., Florentin, D., 2022. Des tuyaux qui comptent. Tournant patrimonial et renégociation des relations entre voirie et réseaux d'eau et d'assainissement, *Flux*, n°128, p. 32-46.

Gouvernement Wallon, 2017. Contrat de gestion entre le gouvernement Wallon et la Société Publique de Gestion de l'Eau, 2017-2022.

Greenwood, M., Yule, G.U., 1920. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. *J. R. Stat. Soc.* 83, 255–279.

Honaker, J., King, G., et Blackwell, M., 2012. AMELIA II: A Program for Missing Data. <https://r.iq.harvard.edu/docs/amelia/amelia.pdf>

Husson, A., Vacelet, A., Stricker, A.E., Bréjoux, E., Renaud, E. 2018. Améliorer la connaissance des réseaux d'eau potable à l'échelle nationale : analyse des données disponibles. *Techniques Sciences Méthodes*, 10, 41-52.

Kabir, H., Solomon, T., Hemsing, J., Sadiq, R., 2020. Handling incomplete and missing data in water network database using imputation methods. *Sustainable and Resilient Infrastructure*, 5:6, 365-377.

Lardon, S., Tonneau, J-P., Raymond, R., Chia, E., Caron, P., 2008. Dispositifs de gouvernance territoriale durable en agriculture. Analyse de trois situations en France et au Brésil, *Norois*, n° 209, p. 17-36.

Lascoumes, P., Le Galès, P., 2018 [2007]. *Sociologie de l'action publique*, Armand Colin.

Le Gat, Y., 2016. Recurrent Event Modeling Based on the Yule Process: Application to Water Network Asset Management. ISTE-Wiley.

Le Gat, Y., 2014. Extending the Yule process to model recurrent pipe failures in water supply networks. *Urban Water J.* 11, 617–630.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models (2nd edition). Monographs on Statistics and Applied Probability 37. Chapman & Hall.

Rahbaralam, M., Modesto, D., Cardus, J., Abdollahi, A., et Cucchiatti, F.M., 2007. Predictive Analytics for Water Asset Management: Machine Learning and Survival Analysis. arXiv:2007.03744.

Rey-Valette, H., Chia, E., Mathé, S., Michel, L., Nougaredes, B., Soulard, C-T., Guiheneuf, P., 2014. Comment analyser la gouvernance territoriale ? Mise à l'épreuve d'une grille de lecture, Géographie, économie, société, n°16, p. 65-89.

Rodriguez, N., Husson, A., Renaud, E., Caillaud, K., Le Gat, Y., Rulleau, B., 2021. Projet GePaME, Rapport intermédiaire 2021.

Rodriguez, N., Husson, A., Le Gat, Y., Renaud, E., 2022. Méthodes automatisées de préparation des données pour la gestion patrimoniale du réseau de la SWDE. 101^e congrès de l'ASTEE, Dunkerque, 14-16 juin 2022.

Ross, S.M., 1996. Stochastic processes, 2nd ed. ed, Wiley series in probability and statistics. Wiley, New York.

SPW, 2017. Traitement de potabilisation des eaux et mise hors service des captages.

13. Annexes

Annexe 1. Comparaison des cellules des conduites recevant les défaillances et des cellules des fichiers fuites sur Mons-Ath-Gaurain (MAG)

Vocabulaire

Cellule de la défaillance = cellule de la conduite = cellule du SIG = numéro de cellule de la conduite du SIG sur laquelle la défaillance a été positionnée.

Cellule des fichiers fuites = cellule « fuite » = cellule inscrite dans les fichiers Excel de fuites pour l'OT correspondant à cette défaillance.

Désaccord = différence entre cellule de la défaillance et cellule des fichiers fuites.

Constat

Les cellules des défaillances sont différentes des cellules des fichiers fuites pour 880 défaillances sur 5 256 (17 %). Ces 5 256 défaillances (concernant conduites ou autres équipements) sont celles pour lesquelles le lien entre l'extrait SAP et fichiers fuites par l'OT est possible sur les secteurs de Mons-Ath-Gaurain (MAG) entre 2016 et 2021 inclus.

Pour rappel, MAG comporte 253 cellules, et parmi ces dernières on observe des désaccords sur 204 cellules.

Objectif

Comprendre pourquoi, et comprendre si on peut utiliser les cellules des fichiers fuites comme critère de sélection lors de l'attribution des défaillances.

Analyses

Tableau 20 : Taux de désaccord entre cellules selon l'année.

Année	Taux de désaccords sur l'année	% du total de désaccords
2016	6.1 %	11 %
2017	5.8 %	13 %
2018	13.9 %	30 %
2019	12.2 %	25 %
2020	7.8 %	17 %
2021	6.4 %	5 %

Il est possible que les taux de désaccords plus élevés en 2018 et 2019 soient dus à la mise en place définitive des nouveaux numéros de cellule intégrant la notion de secteur et au délai de leur prise en compte par les agents de la SWDE.

Tableau 21 : Taux de désaccord entre cellules par secteur.

Secteur	Taux de désaccords sur le secteur	% du total de désaccords
5120 (Gaurain)	19.6 %	45 %

5230 (Mons)	18.9 %	15 %
5110 (Ath)	13.9 %	40 %

Le ratio du total de désaccords semble proportionnel à la taille du secteur (rapport de 3 entre Mons et les autres secteurs). Il est possible que le taux de désaccords spécifique sur Ath soit rendu légèrement plus faible qu'à Gaurain par l'absence de 6 mois de données de fuites Excel de juillet à décembre 2016 inclus. Le taux de désaccords semble au final peu dépendant du secteur.

Les désaccords sont cependant très dépendants des cellules : les 10 cellules de défaillance (i.e. du SIG) ayant le plus de désaccords concentrent à elles seules 30 % des désaccords sur MAG (rappel : 253 cellules dont 204 avec des désaccords).

Tableau 22 : Désaccords de cellules fréquents pour certaines cellules.

Cellule du SIG	Nombre de désaccords	% cumulé du total de désaccords
51108739	70	8
51208944	30	11
52309022	30	15
51208502	23	17
51208501	21	20
51208537	20	22
51108939	18	24
51108730	18	26
51208910	16	28
52300036	15	30

Divisions de cellules

26 désaccords pourraient être expliqués par 5 cellules de défaillance (51108206, 52300036, 52300022, 51108807, 51108711) ayant été divisées au cours de leur vie. Cependant, les cellules en désaccord indiquées dans les fichiers fuites ne correspondent jamais aux cellules filles de la division.

Environ 30 désaccords sont expliqués par 2 cellules des fichiers fuites ayant été divisées. La cellule de la défaillance est alors une des filles de cette division. Il s'agit dans tous les cas sauf 1 de la division 52300022->(52300022+52309022), le « domaine de la brisée » à Obourg (quartier près de Mons dont toutes les rues ont le même ICAR).

Au grand maximum, 100 désaccords pourraient être expliqués par des divisions de cellules (présence d'un numéro ou de l'autre dans la liste des parents ou enfants de division). Ce chiffre est une surestimation certaine.

Renumérotations de cellules

62 désaccords correspondent à une cellule de défaillance qui porte un nouveau numéro suite à une renumérotation. Cependant dans 1 seul cas la cellule de la fuite correspond à l'ancien numéro (52000042->52300042).

62 désaccords correspondent à une cellule de fuite qui porte un nouveau numéro suite à une renumérotation. Cependant dans aucun cas la cellule de la défaillance ne correspond à l'ancien numéro.

Au global, on compte une soixantaine de désaccords qui ne s'expliquent pas par des renumérotations, bien que les cellules concernées soient dans la liste des cellules renumérotées entre 2016 et 2021.

Fautes de frappe

Il y a très peu d'erreurs grossières de frappe immédiatement visibles (numéros aberrants, trop courts...) de numéros de cellule dans les fichiers fuites (moins de 1 %), ce qui est vrai sur MAG comme sur l'ensemble des 17 secteurs de la SWDE.

35 désaccords peuvent être expliqués par 10 numéros de cellule n'existant que dans les fichiers fuites (pas dans SAP et pas dans le SIG, même pas dans les archives SIG) : il s'agit de "51108725" "51208738" "51108231" "51108829" "52030047" "51105642" "51205801" "51108242" "51208938" "523000".

48 % des désaccords entre cellules (426/880) sont limitées à une différence d'un caractère. Et pourtant, une majorité des cellules indiquées de part et d'autre existent. Il y a donc deux possibilités :

- L'erreur de frappe est involontaire mais la cellule indiquée par erreur existe quand même.
- L'erreur de frappe est « volontaire », une cellule adjacente à la cellule de la défaillance est désignée par erreur (peut-être l'habitude de l'agent de pointer les défaillances de cette rue sur cette cellule).

Distances entre cellules en désaccord

581/880 désaccords (66 %) concernent des cellules adjacentes (ou bien sécantes), c'est-à-dire pour lesquelles au moins deux conduites de chaque cellule se touchent ou sont sécantes. Les éléments suivants détaillent les distances entre conduite « choisie » (sélectionnée pour la défaillance) et « cherchée » (la plus proche dans la cellule indiquée dans les fichiers fuites).

Tableau 23 : Distribution des distances (mètres) entre conduite portant la défaillance et la conduite la plus proche de la cellule inscrite dans les fichiers fuites.

min	1 ^{er} quartile	médiane	3 ^{ème} quartile	max	NA
0 m	227 m	671 m	1 646 m	34 885 m	48 cellules des fichiers fuites non présentes dans le SIG 2022.

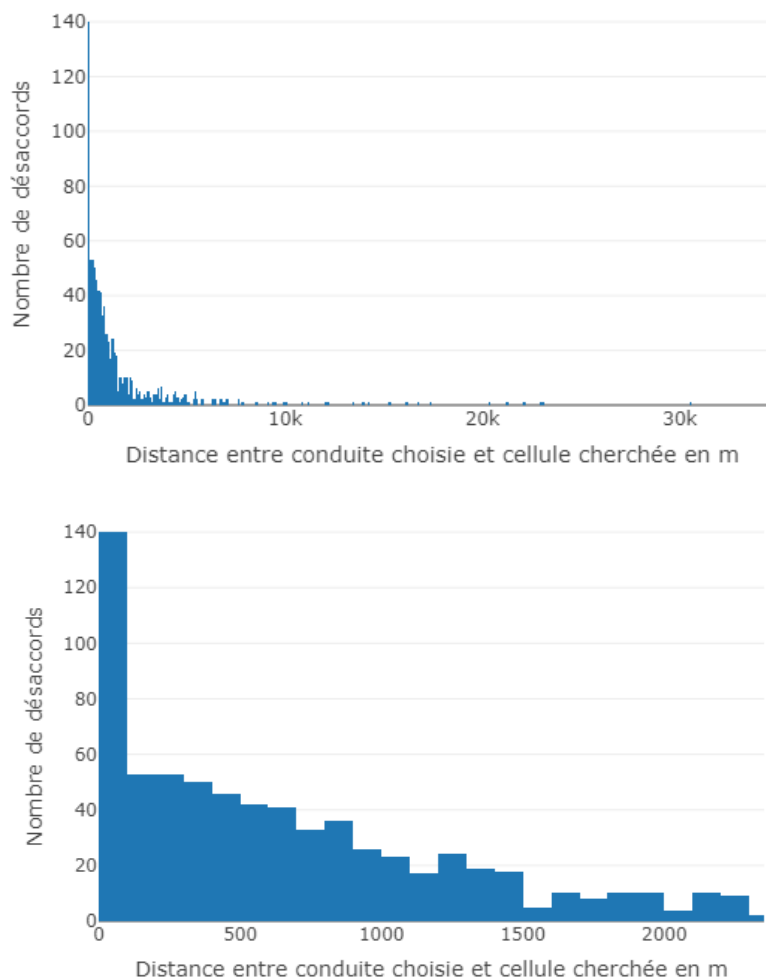


Figure 32 : Distribution des distances (mètres) entre conduite portant la défaillance et la conduite la plus proche de la cellule inscrite dans les fichiers fuites (graphe du bas : zoom).

208 désaccords sont pour des distances inférieures à 227 m entre conduite choisie et cellule inscrite dans les fichiers fuites. Ces désaccords semblent correspondre en ordre de grandeur de distance à des défaillances situées aux interfaces des cellules, et qu'on a affectées à la mauvaise conduite.

119 désaccords sont pour des distances supérieures à 3 km (jusqu'à 35 km), ce qui correspond en ordre de grandeur de distance à des erreurs importantes dans le numéro de cellule inscrit dans les fichiers fuites.

Au final, une large proportion de désaccords (34 %) concerne des cellules qui ne sont pas adjacentes, dont de nombreux cas (plus de 50 % des désaccords) où la cellule des fichiers fuites est loin (plus de quelques centaines de mètres) de la conduite qui porte actuellement la défaillance.

Numéros de cellule des conduites proches de la défaillance

Une analyse des 10 plus proches voisins de chaque conduite défaillante présentant un désaccord avec la cellule des fichiers fuites montre que :

- 612/880 défaillances sont entourées uniquement de conduites de la même cellule.
- Parmi les 268 entourées d'au moins une conduite d'une autre cellule, dans 94 cas il s'agit uniquement de la cellule des fichiers fuites (35 %).
- 812/880 défaillances sont entourées en majorité de conduites de la même cellule.

- 27/880 défaillances sont entourées d'autant de conduites de la même cellule que de cellules indiquées dans « fuites ».
- 41/880 sont entourées de plus de conduites de cellules différentes que de de conduites de la même cellule, dont 19 cas avec seulement la cellule des fichiers fuites.

Ainsi, la majorité des désaccords ne semble pas se situer géographiquement à l'interface de cellules adjacentes, bien que les cellules concernées soient souvent adjacentes.

Influence des attributions de défaillances à la mauvaise conduite

Seulement 60 % des affectations de réparations de conduites à celles du SIG sont sur la bonne conduite (519/867, voir Annexe 4), si l'on se base sur celles qui sont à l'origine localisées sur les conduites du SIG par des « anomalies SIG » ou « opérations SIG » et pour lesquelles cette comparaison est possible. Il est a priori possible que le mauvais positionnement de la défaillance soit à l'origine de changements de cellule et donc du désaccord entre cellule de la défaillance et cellule des fichiers fuites. Cependant, aucune erreur d'affectation parmi les 348 observées ne concerne une autre cellule que la cellule de la « bonne » conduite. Ainsi les erreurs d'affectations des défaillances n'ont a priori aucune chance de provoquer un changement de cellule.

Conclusions

17 % (880/5256) des défaillances affectées aux conduites concernent une cellule que l'on pourrait remettre en question au vu de celle qui est inscrite dans les fichiers fuites. Ces désaccords entre cellules sont présents sur tous les secteurs et toutes les années, mais une proportion importante est spécifique à 2018 et 2019 et à un nombre restreint (quelques dizaines) de cellules. Les divisions de cellule et des renumérotations n'ont pas une influence importante sur ces désaccords. Il y a très peu d'erreurs grossières sur la dénomination des cellules dans les fichiers fuites. Souvent les numéros de cellules en désaccord sont très proches, et concernent donc souvent des cellules adjacentes. Cependant, même lorsque les numéros de cellule sont proches, il peut y avoir une grande distance entre les cellules en désaccord (des dizaines de km parfois). La conduite qui a été choisie est donc plus souvent située au sein de sa cellule qu'à l'interface avec l'autre cellule indiquée. Ainsi la cellule des fichiers fuites n'est pas toujours un critère pertinent de sélection par rapport à la localisation de la défaillance, mais il reste utile.

En définitive, il semble tout de même possible d'utiliser la cellule des fichiers fuites comme critère de sélection d'une conduite pour une défaillance donnée. Cependant, pour éviter les erreurs d'attribution dues uniquement à un numéro de cellule « fuite » imprécis, ce critère ne devra pas être éliminatoire, et la géolocalisation précise de la défaillance (par SIG ou rue avec le numéro d'immeuble) devra primer sur la cellule des fichiers fuites.

Annexe 2. Méthode finale de concaténation des fichiers fuites

Une partie des défaillances du réseau de distribution de la SWDE (principalement les fuites trouvées par la RF) est consignée dans des fichiers Excel (ou onglets) distincts selon le mois ou l'année et selon le secteur, avec des noms et/ou des positions de colonnes parfois différentes pour une même donnée. Pour traiter facilement les données des défaillances sur l'ensemble du territoire de la SWDE, il a été nécessaire de concaténer ces fichiers en un seul en harmonisant au passage toutes les colonnes des fichiers d'entrée. Cela représente plus de 250 fichiers d'entrée à traiter sur les 17 secteurs, ayant chacun en moyenne 35 colonnes à inspecter avant la copie des données. Les colonnes du fichier concaténé représentent bien l'ensemble des colonnes trouvées dans les fichiers de fuites, qui sont les suivantes (Tableau 24).

Tableau 24 : Colonnes du fichier des fuites concaténé

Nom attribué	Signification (parfois interprétée)
annee	année du fichier d'entrée de fuites
num_fuite	numéro attribué à la défaillance par la SWDE
agent1	nom de l'agent (n°1) ayant travaillé sur la localisation et/ou la réparation de la défaillance
commune	
section	correspond aux « zones d'adresse » dans le vocabulaire du Service Public de Wallonie
num_cell	numéro de cellule
nom_cell	nom de cellule
rue	
num_boite	numéro d'immeuble
materiau	matériau de l'élément fuyard
OT_RF	OT « systématique » de recherche de fuite, débute normalement par 7
motif	origine de l'intervention a priori
diametre	diamètre de l'objet fuyard
crash	la défaillance est-elle un crash 0/1
systematique	inverse de crash
OT_reparation	OT de réparation, débute normalement par 5 ou 6
date_localise	date de localisation
par_rf	défaillance trouvée par RF 0/1
par_1ligne	défaillance trouvée par première ligne 0/1
m3_estime	nombre de m ³ perdus estimés
fuite_priv_par_rf	défaillance sur la partie privée (après compteur) trouvée par la RF 0/1
fuite_prod	défaillance sur conduite de production 0/1

permalog	défaillance trouvée via l'utilisation de prélocalisateurs acoustiques permalog 0/1
delai_repa	délai entre localisation et réparation
date_repar	date de réparation
suivi_par	personne en charge du suivi
cause_fuite	cause de défaillance observée a posteriori
qualite_trace	qualité du tracé au sol pour localiser la fuite 0/1
tpr>1.5	est-ce un Trou Pour Rien 0/1 (fuite à plus de 1.5 m du tracé)
fuite_repa_O/N	la défaillance est-elle réparée 0/1
remarque_TPR	remarque sur le Trou Pour Rien
remarque_sect	remarque des agents du secteur
remarque_RF	remarque des agents de RF
TPR_estim_RF	estimation de la pertinence du TPR par les agents de RF
mois	mois de la défaillance
date_loca_2	date de localisation 2 (souvent vide, sinon égal à date_localise)
date_repa_2	date de réparation 2 (souvent vide, sinon égal à date_repar)
m3_perdu_10j	nombre de m ³ perdus estimés sur plus de 10 jours
m3_perdu_100j	nombre de m ³ perdus estimés sur plus de 100 jours
m3/j_10j	débit de pertes estimé sur plus de 10 j
m3/j_100j	débit de pertes estimé sur plus de 100 j
janvier...décembre	12 colonnes précisant la distribution des volumes perdus par mois pour chaque défaillance (souvent vide)
gaz	utilisation de gaz traceur pour trouver la fuite 0/1
permalog_fixe	utilisation de permalog fixe (ENIGMA 3M) 0/1
permalog_mobile	utilisation de permalog mobile 0/1
satellite	utilisation de données satellitaires pour trouver la fuite 0/1
agent2	nom de l'agent (n°2) ayant travaillé sur la localisation et/ou la réparation de la défaillance
secteur	numéro de secteur

Remarque : les données du secteur d'Ath de juillet à décembre 2016 semblent avoir été perdues par la SWDE.

La méthode employée est semi-automatique. La partie manuelle consiste à créer une table de correspondance entre les indices de colonnes des fichiers de données brutes et les indices de colonne du fichier final recevant l'ensemble des données (voir rapport intermédiaire 2021). Il faut pour cela identifier visuellement quelle colonne de départ contient la donnée devant être copiée vers chaque colonne d'arrivée. Heureusement, pour certains secteurs et sur certaines années (voire certains mois) la correspondance est la même et la table de correspondance peut être utilisée d'un secteur à l'autre

ou d'une année à l'autre. L'automatisme (script R) réside simplement dans la lecture et la copie des données ainsi que leur formatage, une fois la correspondance établie. Certains formatages sont importants et sont donc précisés dans la section suivante.

Formatages appliqués aux données

Numéro de cellule

Les numéros de cellule sont normalement formés de 8 chiffres, les 4 premiers étant le code de secteur (tel que 5230 pour Mons) et les 4 derniers étant le code de la cellule pour ce secteur (tel que 0004 pour la cellule n°4). Le format des numéros de cellule peut différer de la règle ci-dessus d'un secteur à l'autre (anciens noms alphanumériques, numéro à 5 chiffres, trait d'union entre secteur et cellule...). Pour uniformiser les numéros de cellule, la procédure suivante a été appliquée :

- i. Suppression du trait d'union éventuellement présent entre code secteur et code cellule (aperçu sur certains secteurs).
- ii. Extraction de tout nombre ayant 1 à 8 chiffres (inclus) comme étant le numéro de cellule.
- iii. Pour tout nombre extrait ayant seulement 1 à 4 chiffres (inclus), ajout du code secteur.

Ainsi, tout numéro de cellule extrait ayant entre 5 et 7 chiffres inclus est conservé tel quel, mais probablement erroné.

Extraction des diamètres

Les diamètres sont extraits en premier lieu du champ « diamètre » si ce dernier est présent (généralement à partir de l'année 2018, bien que certains secteurs ne le remplissent pas toujours), en cherchant une expression régulière de type *Chiffre* (par exemple 80 pour le format en diamètre nominal) ou *ChiffrePonctuationChiffre* (par exemple 3/4 pour les diamètres en format NPS).

Dans un second temps, pour les diamètres n'ayant pas pu être récupérés par l'étape précédente, le diamètre est extrait du champ « matériau » en recherchant la même expression régulière que précédemment (entre 2016 et 2018 le diamètre est plutôt renseigné dans la même colonne que le matériau, par exemple « FONTE DN 80 » ou « 80 fonte »).

Remarque : la conversion de NPS à DN n'est pas effectuée ici. Elle est faite seulement lors de l'affectation des défaillances aux conduites, qui utilise le fichier de fuites concaténé.

Formatage des matériaux

Les matériaux présentent une grande hétérogénéité d'écritures (par exemple pour amiante ciment : asb, a.c., ac, iso, ét, eb, amiante...). Un ensemble d'expressions régulières a été utilisé pour reconnaître et renommer les matériaux, selon les règles suivantes (Tableau 25).

Tableau 25 : Expressions régulières utilisées pour extraire les matériaux.

Expression régulière détectée dans le champ « matériau »	Matériau affecté
<code>fte fonte dn fonte de fonte[:blank:]*[:digit:]+ fonte[:blank:]*\$ ^[:blank:]*f[:blank:]*[:digit:]+ [:digit:]+[:blank:]*f\$</code>	FONTE
<code>f[:blank:]*g fonte[:blank:]*grise</code>	FONTE GRISE
<code>f[:blank:]*d fonte[:blank:]*ductile</code>	FONTE DUCTILE
<code>pvc</code>	PVC
<code>acier ac[:blank:]*[:digit:]+ ac dn ac[:blank:]*\$ ^[:blank:]*a[:blank:]*[:digit:]+ [:digit:]+[:blank:]*a\$</code>	ACIER
<code>asb a\\.c\\. amiante et eb ét iso</code>	AMIANTE CIMENT
<code>bonna</code>	BETON AME TOLE
<code>plomb pb</code>	PLOMB
<code>pe poly</code>	PE

La syntaxe des expressions régulières est expliquée dans la « Cheat Sheet » de la librairie `stringr` de R (utilisée pour les traitements de texte).

Remarque : le matériau FONTE sans précision est assimilé à de la FONTE GRISE lors de l'import du fichier fuites concaténé pour l'affectation des défaillances aux tronçons du SIG. En effet, le SIG ne contient pas de conduites en FONTE, seulement de la FONTE GRISE ou de la FONTE DUCTILE. Il faut donc faire un choix. Ici nous partons du principe que le matériau le plus probable (car le plus représenté sur le réseau) est la FONTE GRISE.

Extraction des OT et OT RF

Les deux colonnes « OT » qu'on peut trouver dans les fichiers bruts contiennent parfois des OT de réparation, parfois des OT de RF, parfois un mélange des deux (par exemple « 5XXXXXXXX/7XXXXXXXX »). Ceci peut arriver même lorsque deux colonnes OT ont été créées explicitement pour distinguer les OT de réparation des OT systématiques (à partir de 2018, les OT RF ne sont généralement plus renseignés). Afin de toujours affecter le bon type d'OT à chaque colonne du fichier concaténé (soit OT_reparation, soit OT_RF), une reconnaissance du premier chiffre de l'OT a été effectuée de façon automatique :

- 1) Pour chaque défaillance, si OT_reparation est vide, on tente de le retrouver en lisant dans la colonne OT_RF un numéro à 9 chiffres commençant par un 5 ou un 6.
- 2) Pour chaque défaillance, si OT_RF est vide, on tente de le retrouver en lisant dans la colonne OT_reparation un numéro à 9 chiffres commençant par un 7.

Les étapes (1) et (2) permettent ainsi de régler le problème des mélanges entre colonnes.

- 3) Pour chaque défaillance, conserver dans OT_reparation uniquement un numéro à 9 chiffres commençant par un 5 ou un 6 parmi le contenu actuel de la colonne.

4) Pour chaque défaillance, conserver dans OT_RF uniquement un numéro à 9 chiffres commençant par un 7 parmi le contenu actuel de la colonne.

Les étapes (3) et (4) effectuent donc un nettoyage des colonnes OT_reparation et OT_RF en supprimant tout ce qui n'est pas dans le format attendu.

Suppression des doublons d'OT

Certains secteurs répètent les mêmes OT d'année en année (parfois sur plusieurs années). Ceci est probablement dû à une mise à jour tardive des informations (défaillances de décembre réparées en janvier suivant) ou à des réparations de défaillance effectuées tardivement après leur localisation. Afin d'éviter d'avoir des doublons d'OT avec des informations contradictoires dans le fichier final, une sélection des informations a été appliquée : pour tout OT répété, une seule ligne est gardée, et l'information conservée pour chaque colonne (par exemple le matériau) correspond à la dernière apparition chronologique de l'OT où l'information cherchée est renseignée. Ce filtre a été appliqué seulement sur les informations utilisées pour l'affectation des défaillances aux tronçons du SIG, il s'agit des attributs : commune, rue, num_boite, materiau, diametre, crash, par_rf, date_localise, date_repar. Pour les autres colonnes, c'est la valeur (vide ou non) de la dernière ligne qui est choisie.

Formatage des dates

Les dates de localisation et de réparation peuvent également présenter des formats différents : 01 janvier 2006, 01/01/2006, 01/01/06, ou encore 01-01-2006. L'homogénéisation a été faite dans Excel, par une conversion des données de la colonne vers le type date, et par un formatage en date de type j/m/a. Lors de l'import et de la lecture du fichier dans R, le motif cherché est donc j/m/a.

Annexe 3. : Statut crash et détection par RF des défaillances

Une défaillance a été considérée comme un crash si (conditions suffisantes) :

- L'OT est associé à une « OPERATION SIG ».
- L'OT est associé à « crash=1 » dans les fichiers Excel de fuites.
- L'OT est de type CLT et date de 2020 ou 2021.
- L'OT n'est pas dans les fichiers Excel de fuites entre 01/01/2016 et 31/12/2019.

Les « opérations SIG » sont majoritairement des défaillances de conduite réparées dans l'urgence sans anomalie SIG constatée au préalable. Une analyse croisée (présentée au CSO du 08/12/2021) sur la période 2007-2021 du type d'ordre, de la présence dans le SIG de l'OT, et du statut crash O/N dans les fichiers fuites pour les réparations de conduite sur les trois secteurs de Mons, Ath, et Gaurain a révélé que seulement à partir de 2020 inclus, le type d'ordre permet clairement de distinguer les crashes des non-crashes. Enfin, le dernier point provient d'échanges avec la SWDE révélant que tout OT qui n'est pas dans les fichiers Excel est normalement un crash.

Les défaillances à partir du 01/01/2016 qui ne remplissent aucunes des conditions suffisantes ci-dessus ont été considérées comme n'étant pas des crashes a priori. Les défaillances avant 2016 ont reçu le statut crash « inconnu », par manque d'informations permettant de tester les conditions ci-dessus.

Un marqueur a été ajouté aux défaillances pour identifier l'origine de la donnée qui a permis d'identifier qu'il s'agit d'un crash le cas échéant : OpeSIG/HorsXL/TA_SAP/CrashExcel.

Une défaillance a été considérée comme trouvée par la RF si (condition suffisante) :

- L'OT est dans les fichiers fuites et l'attribut « par RF » vaut 1.

Toutes les défaillances qui ne remplissent pas la condition ci-dessus ont donc été considérées a priori comme trouvées autrement que par la RF. De plus, toutes les défaillances avant 2016 (qui ne sont donc pas dans les fichiers fuites) ont reçu une méthode de détection « inconnue », par manque d'informations.

Annexe 4. Validation de la méthode d'affectation des défaillances de conduites aux tronçons du SIG

Introduction

Dans le Système d'Information (SI) de la SWDE, une proportion importante des défaillances (sur canalisations, raccordements, bouches d'incendie, etc.) ne sont pas localisées précisément par leurs coordonnées géographiques dans le SIG, mais seulement par le nom de la rue où elles se produisent, avec parfois la précision du numéro de l'immeuble. Par exemple, un peu moins de 20 % des réparations de conduite (entre 2007 et 2021 sur les secteurs de Ath, Gaurain, et Mons) sont précisément géolocalisées dans le SIG, sur la conduite défaillante elle-même. Si la localisation des défaillances dans le SIG permet de les affecter de façon certaine à un tronçon de réseau, il n'en va pas de même de la localisation « floue » par rue, ou couple rue - numéro d'immeuble. Un gros effort de mise au point algorithmique (voir rapport intermédiaire 2021) a donc été consenti dans le cadre de la tâche 01 du projet GePaME, afin d'établir une procédure d'exploitation de l'ensemble des informations disponibles sur les tronçons de réseau et sur la réparation des défaillances permettant d'affecter ces dernières aux tronçons les plus probables. De la réussite de ce travail dépend nécessairement le bon déroulement des tâches 03 (intégration du logiciel Casses dans le SI), 04 (intégration des modules d'aide à la décision dans le SI), et plus largement de l'ensemble des autres tâches concourant à l'élaboration des outils d'aide à la décision.

Objectifs de la validation

La procédure d'affectation des défaillances non géolocalisées aux tronçons est une combinaison complexe de traitements de données géographiques et de documentation des tronçons et des réparations de défaillances. Son adoption par des opérateurs requiert que sa fiabilité soit mise à l'épreuve par un travail de validation. A cet effet, un jeu de défaillances dont l'affectation est sûre a été dégradé en ignorant l'information SIG, et ré-affecté aux tronçons en utilisant les seules informations de nom de rue ou couple rue - numéro d'immeuble disponibles. Il en résulte un jeu de défaillances permettant de comparer la « bonne » affectation (via le SIG), à celle résultant de la procédure à valider. L'objet de ce travail n'est pas la validation complète de la procédure, mais d'apprécier l'impact des erreurs (inévitables) d'affectation, notamment sur le phénomène d'agrégation tendancielle des défaillances sur les mêmes tronçons; il est en effet important, dans la perspective de modélisation LEYP du processus de défaillance, ainsi que d'étude de la probabilité d'état fuyard des tronçons, que la procédure d'affectation perturbe le moins possible la nature agrégative du phénomène de défaillance.

Méthode employée

Ce travail de vérification se focalise sur les défaillances de type « réparations de conduite » (défaillances de type d'activité 105 dans SAP), qui représentent donc les conduites qui ont eu des fuites ou des crashes. Le jeu de données utilisé pour cet exercice consiste ainsi en 867 réparations de conduite enregistrées sur les secteurs de Mons, Ath, et Gaurain entre 2007 et 2021, qui avaient été précisément affectées aux conduites dans le SIG par géolocalisation, et qui ont pu être réaffectées aux conduites sans la géolocalisation du SIG. Il est à noter que les défaillances de conduite qui n'ont pas pu être réaffectées sans la géolocalisation précise SIG sont ignorées (ceci se produit lorsque les informations de localisation disponibles sont insuffisantes ou contradictoires avec le matériau ou le diamètre renseigné pour cette défaillance).

Le taux de succès de l'affectation des défaillances de conduite selon la méthode et la disponibilité des données de localisation est étudié :

- En faisant un simple ratio du nombre de succès par le nombre de cas.
- Par une régression logistique (avec fonction de lien logit).

De plus, les caractéristiques des conduites (appelées covariables dans l'optique du LEYP) et la rue dans laquelle se situent les conduites sont également comparées après réaffectation des défaillances.

Résultats

La comparaison des conduites recevant les défaillances révèle que :

- 60 % des défaillances de conduite sont réaffectées à la même conduite.
- 76 % des défaillances de conduite sont affectées à une conduite (potentiellement la même) qui possède des covariables (matériau, diamètre, date de pose) identiques.
- 17 % des défaillances de conduite sont réaffectées sur une autre conduite, mais qui possède les mêmes covariables que la conduite qui a été défaillante (près de 42 % des attributions erronées sont de ce type).
- 89 % des défaillances de conduite sont affectées sur une conduite (potentiellement la même) située dans la même rue.
- L'incertitude principale d'attribution est la localisation ambiguë de la défaillance.
- En cas d'affectation incorrecte, la conduite choisie est souvent proche de la conduite défaillante.
- Les conduites mal choisies par la rue seule sont en moyenne plus longues (impact clair du choix de prendre la conduite la plus longue de la rue à défaut d'autres informations).
- Les erreurs d'affectation ne semblent pas impacter le phénomène de défaillances répétées, et provoquent seulement une légère concentration sur certaines conduites, qui peut être négligée.
- Le taux d'erreur d'affectation dépend fortement de la disponibilité des données : la présence d'un numéro d'immeuble et/ou de renseignements sur le matériau et/ou le diamètre de la conduite augmentent le taux d'affectations réussies, ce qui est résumé dans le tableau ci-dessous.

Tableau 26 : Taux de succès d'affectation par méthode.

Informations disponibles		Nombre de défaillances de conduites	Nombre de succès	Taux de succès observé	Taux de succès prédit	Taux de covariables égales	Taux dans la même rue
Numéro d'immeuble	Données MAT ou DIA						
NON	NON	197	55	0.28	0.26	0.53	0.89
NON	OUI	82	35	0.43	0.46	0.76	0.85
OUI	NON	178	104	0.58	0.60	0.69	0.79
OUI	OUI	410	325	0.79	0.79	0.91	0.95
TOTAL		867	519	0.60	-	0.76	0.89

Annexe 5. : Corrections topologiques sur les conduites du SIG

Conduites en service

Fichier : Conduites en service du 11/02/2022 concernant le périmètre SWDE entier.

Tableau 27 : Conduites ES corrigées.

Conduite concernée	Problème	Action effectuée
AQS_16415240	Plusieurs morceaux non connexes	Regroupement en suivant le tracé de la conduite STARAPIC-414736 (caractéristiques identiques) qui est donc supprimée (elle ne porte pas d'anomalie SIG et fait 2 m).
AQS_16415124	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
AQS_16415129	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant deux nœud du plus petit
AQS_16414675	Morceau doublon isolé	Suppression du morceau inutile
AQS_16413051	Morceau doublon isolé	Suppression du morceau inutile
AQS_16414401	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
AQS_16416497	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
200033818	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
AQS_16414727	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
600008779	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
600006094	Morceau doublon isolé	Suppression du morceau inutile
600005938	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit

500014748	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
500018374	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
500015508	Morceau doublon isolé	Suppression du morceau inutile
500015509	Nœud intermédiaire faux créant un angle très aigu	Suppression du nœud intermédiaire
600018265	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
STARAPIC-424856	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
STARAPIC-415127	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
STARAPIC-407886	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
STARAPIC-422173	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
mav_16441567	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
sam_16435211	Nœud intermédiaire faux créant une auto-intersection	Suppression du nœud intermédiaire inutile
STARAPIC-11340	Tronçon replié sur lui-même	Suppression d'un des nœuds non accrochés à un autre tronçon
STARAPIC-363148	Conduite superposée à elle-même (en V)	Suppression d'un nœud doublon
STARAPIC-420109	Auto-intersection	Décalage d'un nœud terminal de 9 mm pour éviter l'auto-intersection
STARAPIC-394746	Nœud intermédiaire faux créant une auto-intersection	Suppression du nœud intermédiaire inutile

400023703	Conduite raccourcie et doublon de STARAPIC-65777 (caractéristiques identiques)	Suppression de 400023703
700005338	Nœud intermédiaire faux créant une auto-intersection	Suppression du nœud intermédiaire inutile
800012441	Nœud intermédiaire faux créant une auto-intersection	Suppression du nœud intermédiaire inutile
800000649	Nœud intermédiaire faux créant une auto-intersection	Suppression du nœud intermédiaire inutile
mav_16442361	Conduite superposée à elle-même (en V)	Suppression d'un nœud doublon
500026401	Conduite dans une chambre à surpresseur formant une boucle rectangulaire étrange	Pas de modification apportée car probabilité d'erreur de dessin faible
STARAPIC-384929	Conduite dans une chambre à surpresseur formant une boucle rectangulaire étrange	Pas de modification apportée car probabilité d'erreur de dessin faible
STARAPIC-311284	Conduite formant une boucle rectangulaire pour desservir un quartier	Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon
STARAPIC-329239	Conduite formant une boucle pour desservir un quartier	Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon
STARAPIC-421879	Conduite formant une boucle sur elle-même	Suppression de deux nœuds inutiles
400004548	Superposition avec 400004549	Suppression de 400004549 qui est en DN60 alors que le DN100 paraît plus logique.
STARAPIC-307758	Superposition avec STARAPIC-307761	Conduites doublons, suppression de STARAPIC-307758
STARAPIC-399600	Superposition avec STARAPIC-399596	Conduites doublons, suppression de STARAPIC-399596
MANU_C_000321	Superposition avec MANU_C_000034	Suppression de MANU_C_000034 : DDP moins cohérente

STARAPIC-347340	Superposition avec STARAPIC-347339	Conduites doublons, suppression de STARAPIC-347339
200000932	Superposition avec 200000933	Suppression de 200000932 qui a un DN moins cohérent
200005178	Superposition avec 200005170	Suppression de 200005178 qui a un DN moins cohérent
STARAPIC-303306	Superposition avec 400012464	Suppression de 400012464 qui a un DN moins cohérent avec le té DN90
400015299	Superposition avec 400015297	Suppression de 400015299 qui est moins cohérent
MANU_C_003084	Superposition avec AQS1_16414153	Conduites doublons, suppression de MANU_C_003084
100013821	Superposition avec STARAPIC-181705	Suppression de 100013821 qui paraît moins cohérente
STARAPIC-318343	Superposition avec STARAPIC-318218	Conduites doublons, suppression de STARAPIC-318218
STARAPIC-406826	Superposition avec STARAPIC-358464	Conduites doublons, suppression de STARAPIC-358464
STARAPIC-318342	Superposition avec STARAPIC-318339	Conduites doublons, suppression de STARAPIC-318339
STARAPIC-318344	Superposition avec STARAPIC-318341	Conduites doublons, suppression de STARAPIC-318341
Toutes		st_line_merge

En résumé : 31 corrections ont été apportées sans modification du nombre ou du tracé des conduites, 17 conduites doublons ont été supprimées, et 4 conduites à la topologie étrange ont été conservées. Cela représente environ 3 km de conduites traitées au total.

Suite à ces corrections, une fusion des polygones connexes (st_line_merge) a été effectuée dans R pour fusionner les éventuels morceaux multiples mais connexes d'une même conduite. La vérification topologique sous QGIS mais aussi sous R n'a par la suite plus fait ressortir d'autres erreurs (hormis les 4 conduites à la topologie étrange mais ignorées).

Conduites Hors Service

Fichier : Conduites abandonnées en date du 11/02/2022 sur le périmètre SWDE entier.

Note importante : suite au grand nombre de conduites HS superposées imposant une suppression automatisée des doublons, le script R de détection des conduites superposées développé pour Mons a été modifié pour identifier les ID des doublons à supprimer comme étant en premier lieu ceux des conduites qui ont le moins de caractéristiques renseignées du groupe de conduites superposées (caractéristiques utilisées sans ordre de priorité : NATURE, TYPE, ABANDON, ANNEE). En second lieu, lorsque les doublons restants ont le même nombre de caractéristiques renseignées, un seul est conservé aléatoirement sans étape de vérification de la plus ou moins bonne cohérence de ses données par rapport aux conduites adjacentes et aux autres doublons.

Ainsi, les conduites HS superposées sont supprimées en masse, avec une identification automatique des ID à supprimer basée seulement sur le nombre de caractéristiques renseignées, contrairement aux conduites ES où le traitement est fait à la main en inspectant le SIG et en déterminant quelle conduite est la plus cohérente à garder.

Tableau 28 : Conduites HS corrigées.

Conduite HS concernée (ID)	Problème	Action effectuée
9386	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
9402	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
9405	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
28942	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
31370	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
32908	Deux cassures au milieu	Attachement des trois morceaux
24886	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
24888	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit
24914	Plusieurs morceaux non connexes	Attachement des morceaux en déplaçant un nœud du plus petit

30545	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
30553	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
26163	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
26166	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
26168	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
31969	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
31970	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
27161	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
29823	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
37533	Deux cassures au milieu			Attachement des trois morceaux
37532	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
46703	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
46706	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit
46380	Plusieurs connexes	morceaux	non	Attachement des morceaux en déplaçant un nœud du plus petit

44341	Plusieurs morceaux connexes	non	Attachement des morceaux en déplaçant un nœud du plus petit
44347	Plusieurs morceaux connexes	non	Attachement des morceaux en déplaçant un nœud du plus petit
45737	Plusieurs morceaux connexes	non	Attachement des morceaux en déplaçant un nœud du plus petit
46288	Plusieurs morceaux connexes	non	Attachement des morceaux en déplaçant un nœud du plus petit
46289	Plusieurs morceaux connexes	non	Attachement des morceaux en déplaçant un nœud du plus petit
46298	Plusieurs morceaux connexes	non	Attachement des morceaux en déplaçant un nœud du plus petit
1855	Conduite formant une boucle rectangulaire		Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon
3610	Conduite formant une boucle rectangulaire		Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon
3617	Conduite formant une boucle rectangulaire		Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon
4518	Conduite formant une boucle rectangulaire		Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon
4538	Conduite formant une boucle rectangulaire		Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon
21462	Conduite formant une boucle rectangulaire		Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon

18813	Conduite formant une boucle rectangulaire	Pas de modification apportée car probabilité d'erreur de dessin faible et pour éviter la création d'un nouveau tronçon
Suppression automatique de 206 doublons de conduites superposées (16 km)		
Toutes		st_line_merge
Suppression de 545 conduites de longueur < 0.5 m qui engendreraient des conduites réduites à un point après accrochage appliqué ci-dessous		
Toutes		Accrochage des nœuds terminaux distants de <= 0.5 m
Corrections finales suite à des nouvelles erreurs topologiques dues à l'accrochage :		
31796	Réduit à un point suite à l'accrochage	Suppression
4667	Erreur de dessin	Suppression
		Suppression manuelle de 8 doublons :
2161	Superposé à 2169	Supprimer et garder 2169 qui est mieux renseigné
2162	Superposé à 2166	Supprimer et garder 2166 qui est mieux renseigné
33481	Superposé à 33477	Supprimer et garder 33477 qui est mieux renseigné
2304	Superposé à 2301	Doublons identiques, supprimer 2301
4442	Superposé à 3343	Supprimer 4442 qui paraît moins cohérent
30325	Superposé à 15721	Supprimer 30325 qui paraît moins cohérent
6193	Superposé à 6184	Suppression des deux qui semblent être une erreur de dessin
23651	Superposé à 23683	Doublons identiques, supprimer 23683

Les conduites HS traitées représentent environ 20 km de réseau HS : 4 km de traitement manuel et 16 km de traitement automatique.

Annexe 6. Étude de l'impact de la complétion de données des conduites sur les prédictions de défaillances avec le LEYP

Revue et amélioration des méthodes de complétion existantes

Recherche bibliographique

La première étape a consisté à faire une revue bibliographique des méthodes de complétion existantes sur les réseaux AEP ou autres réseaux enterrés (comme les réseaux d'assainissement, si la méthode est transférable aux réseaux AEP). L'exploration bibliographique a néanmoins révélé qu'il existe très peu de méthodes publiées dans ce domaine, avec moins de 5 articles considérés comme traitant précisément de la problématique (Belghaddar et al., 2021 ; Betrie et al., 2016 ; Honaker et al., 2012 ; Kabir et al., 2017 ; Rahbaralam et al., 2020).

Essentiellement, les méthodes proposées dans la littérature explorée ne considèrent que rarement la localisation et l'adjacence des conduites, et ignorent donc le fait que des conduites adjacentes ont probablement beaucoup de caractéristiques en commun, contrairement à la méthode qui a été employée pour les conduites HS de la SWDE. Ont donc été publiées essentiellement des méthodes qui estiment les valeurs manquantes par groupes de conduites non connexes ayant une ou plusieurs caractéristiques en commun. L'exemple le plus simple consiste à prendre la médiane de la DDP de toutes les conduites en fonte grise, pour toutes les conduites en fonte grise dont la DDP n'est pas renseignée. Néanmoins, la littérature fournit quelques exemples utiles d'indicateurs de qualité et/ou de quantité de la complétion, qui ont servi d'inspiration aux indicateurs utilisés par la suite (voir section « Indicateurs »).

Méthodes de complétion sélectionnées

Au final, les méthodes suivantes ont été choisies pour comparaison :

- m1 : méthode médiane ; il s'agit de prendre la médiane (ou le mode, si la variable est qualitative) de la donnée manquante pour toutes les conduites qui ont la même modalité pour une autre caractéristique (par exemple le matériau si l'on souhaite compléter la date de pose ou le diamètre si on cherche à compléter le matériau).
- c1 : méthode de complétion de proche en proche détaillée dans le rapport intermédiaire de 2021. Ici, le rapprochement entre nœuds des réseaux ES et HS n'est pas considéré, il s'agit uniquement de la phase de propagation au sein d'un même réseau connexe (ici le réseau ES).
- c2 : amélioration de la méthode c1 avec des règles « métier » ; il s'agit d'interdire la propagation de proche en proche si une ou plusieurs caractéristiques des conduites adjacentes considérées ne coïncident pas (voir Figure 33). Il est par exemple possible de propager les DDP sous condition que les conduites adjacentes aient le même matériau et le même diamètre, ou bien les matériaux sous condition que les conduites adjacentes aient seulement le même diamètre.
- c3 : amélioration de c1 pour contourner les blocages ; il s'agit d'appliquer la méthode c1 de façon répétée sur le réseau, en utilisant chaque matériau séparément, et de stocker les résultats de la complétion par matériau dans des attributs temporaires. Par exemple, pour la DDP, chaque conduite reçoit autant de DDP temporaires qu'il y a de matériaux dans le réseau connexe à cette conduite. L'attribut récupéré correspond au final à l'attribut temporaire reçu de la part du même matériau que la conduite n'ayant pas l'attribut. Les conduites reçoivent donc l'information manquante de la part de la conduite de même matériau la plus proche ayant cette information. Cette méthode permet d'avoir une règle métier comme c2, et en plus d'éviter qu'un morceau de tuyau d'un matériau différent de celui d'une rue entière bloque la propagation de l'information dans cette rue. La méthode c3 peut également être employée avec d'autres caractéristiques que

le matériau conditionnant la complétion. La caractéristique choisie ne doit pas posséder trop de modalités afin de limiter le temps de calcul (n modalités donnent en effet lieu à n complétions indépendantes).

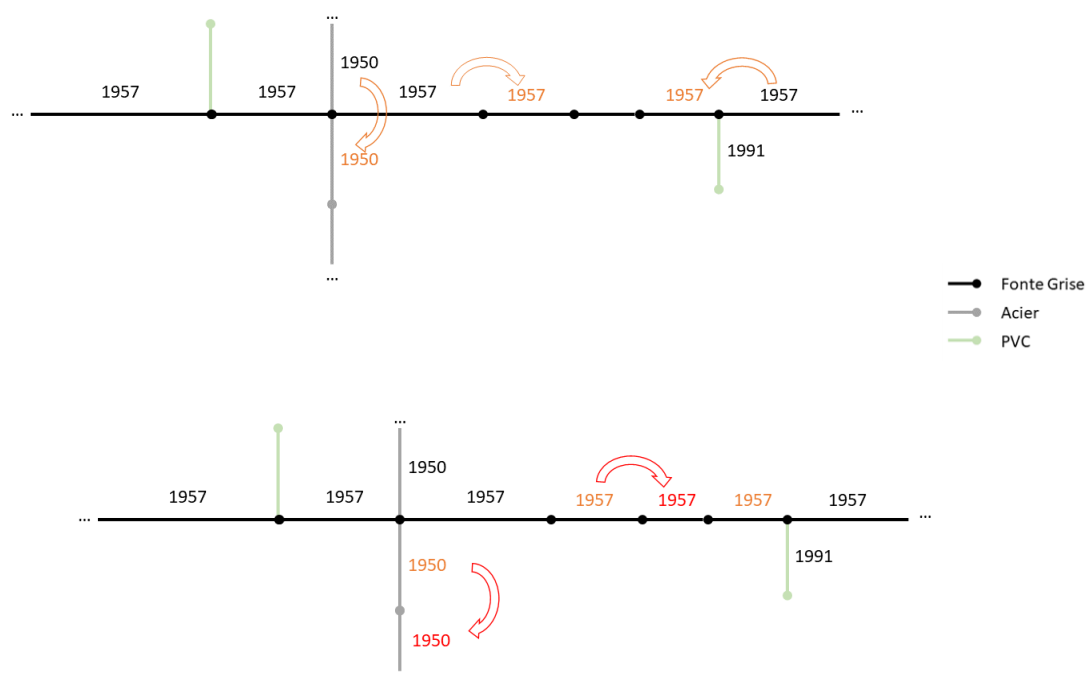


Figure 33: Illustration de la complétion de DDP en fonction du matériau par la méthode c2 sur un réseau fictif

Validation des méthodes de complétion

Objectifs

Le but de la validation est de comparer la performance des 4 méthodes sélectionnées, définie par deux critères simples :

- Taux de vides complétés (« quantité »).
- Taux d'informations correctement complétées (« qualité »).

Ainsi, il ne s'agit pas uniquement de compléter au maximum les données, mais également de compléter par l'information la plus pertinente. Par exemple, cela n'a pas beaucoup de sens de donner une DDP de 1950 (provenant d'une fonte grise) à des conduites en fonte ductile. Concernant le critère de qualité, il a été impératif de partir d'un jeu de données où l'information d'origine est fiable et connue en grande proportion, afin de pouvoir artificiellement oublier cette information, la compléter par les méthodes ci-dessus, et comparer les résultats complétés aux originaux. Ainsi, il a été nécessaire de se restreindre au réseau en service où très peu d'attributs des conduites sont manquants. De plus, ce réseau présente l'avantage d'être connexe (contrairement au réseau HS), ce qui est essentiel pour le bon fonctionnement des méthodes c1, c2, et c3. Sans la connexité, il aurait été nécessaire de chercher l'information autour des paquets de conduites, de la même façon que cela a été fait pour le réseau HS.

Choix de l'attribut à compléter

Afin de ne pas avoir d'interactions trop importantes pendant la complétion entre les divers attributs pouvant être complétés, et par la suite lors des calages LEYP, il a été choisi de se concentrer dans un premier temps sur une seule variable à compléter, en particulier la DDP. Ce choix correspond également à la donnée qui possède le plus de vides à compléter sur le réseau de la SWDE (en particulier HS).

Méthodes de génération des oublis de données

Deux méthodes d'oubli des données ont été retenues :

- Oubli totalement aléatoire : un échantillon aléatoire de conduites sur l'ensemble (par exemple 40 % du nombre, qui peut être converti en % de la longueur totale) dont l'attribut (ici DDP) est connu perdent temporairement l'information. La distribution spatiale du manque de données est donc uniforme.
- Oubli par ICAR de rue : un échantillon aléatoire de rues est sélectionné (par exemple 10 % des rues en nombre). Toutes les conduites dans ces rues perdent temporairement l'attribut. Cette méthode possède un caractère plus réaliste que l'oubli aléatoire, car la distribution spatiale du manque de données n'est plus uniforme, mais forme des petits groupes (rues seules ou groupes de rues adjacentes).

On peut noter que d'autres méthodes d'oubli pourraient être testées si plus de temps était disponible (oubli par commune, ce qui n'est pas envisageable pour un secteur donné, qui ne possède que quelques communes; oubli par âge de conduite, en partant du principe par exemple que les DDP les plus anciennes ont été plus fréquemment oubliées).

Indicateurs

Les indicateurs de performance de complétion utilisés sont les suivants :

- Quantité (taux) de la complétion :
 - I_{LC} : longueur de conduites complétées / longueur de conduites à compléter.
- Qualité (taux d'erreurs) de la complétion :
 - I_{LE} : longueur de conduites avec erreurs de complétion (donnée complétée \neq donnée d'origine) / longueur de conduites à compléter.

Résultats de la validation

La validation a été effectuée sur le secteur de Mons en utilisant la DDP avec des % d'oubli de donnée croissants (Figure 34, oubli par ICAR de rue). Pour la méthode c2, la complétion de la DDP a été conditionnée par la présence du même matériau. La propagation de l'information a été limitée aux 20 conduites adjacentes (limitant le temps de calcul et garantissant la cohérence de la complétion). À chaque % d'oubli de données considéré (5 %, 10 %, 15 %, ...), 10 essais ont été réalisés (Monte Carlo) afin de prendre en compte le caractère aléatoire de l'oubli et d'éviter les conclusions particulières dues purement au hasard.

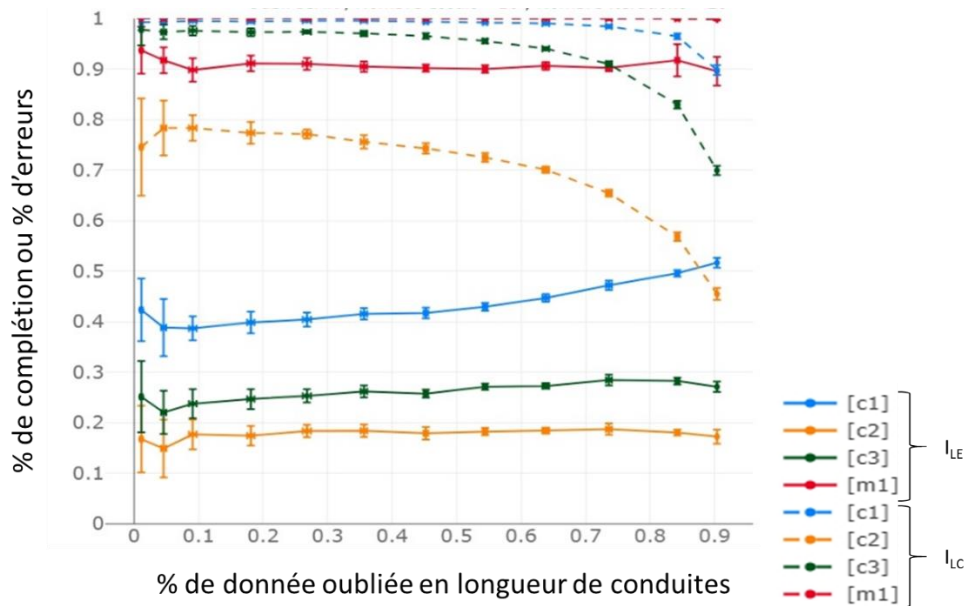


Figure 34 : Performance de la complétion de DDP sur le secteur de Mons (oubli par ICAR de rue).

Cette comparaison montre clairement un nombre très important d'erreurs de complétion (proche de 95 %) avec la méthode m1. Néanmoins, l'analyse des erreurs révèle qu'en moyenne ces dernières sont de plus faible amplitude en DDP avec la méthode m1. Moins d'erreurs sont observées avec c2 que c1 ($I_{LE}(c2) < I_{LE}(c1)$), ce qui est attendu. Néanmoins, contrairement à l'application de c1, il n'est pas possible de compléter la quasi-totalité des données manquantes avec la méthode c2 ($I_{LC}(c2) < I_{LC}(c1)$). Un bon compromis entre c1 et c2 est atteint avec la méthode c3, permettant d'obtenir un taux de complétion élevé et un taux d'erreur de complétion maîtrisé. Utiliser c3 possède cependant l'inconvénient d'être très coûteux en temps de calcul, ce qui est pénalisant à cause de la nécessité d'utiliser de nombreux essais pour limiter le caractère aléatoire des résultats.

Des conclusions similaires peuvent être tirées de la comparaison des méthodes sur Mons en oubliant les données de façon totalement aléatoire, et/ou en inspectant les taux de complétion et d'erreur par rapport aux nombres de conduites au lieu de leurs longueurs. Oublier les données de façon aléatoire permet d'obtenir des taux d'erreurs I_{LE} environ 2 fois plus faibles qu'avec un oubli par ICAR de rue. Cependant, un oubli aléatoire des données n'est pas représentatif de la distribution réelle des manques de données sur un réseau d'eau potable.

En conclusion, chacune des méthodes comparées présente des caractéristiques intéressantes (telles que les erreurs de complétion pour m1, qui pourra représenter les données « mal complétées »). Toutes les méthodes ont donc été conservées pour vérifier leur effet sur la prédiction des défaillances, hormis la méthode c3 qui est trop coûteuse en temps de calcul.

Résultats sur les prédictions de défaillances

Mesurer l'impact de la complétion de données sur les prédictions de défaillances nécessite de se pencher sur les questions ci-dessous, dont les réponses sont apportées dans les sous-sections suivantes :

- Quelles données sont utilisées parmi tous les secteurs de la SWDE.
- Quel attribut des conduites est oublié puis complété, et de quelle(s) manière(s).

- Comment sont comparés les calages des modèles utilisant les différents jeux de données résultants du point précédent.

Choix du secteur de référence

La première étape a consisté à choisir un secteur de référence sur lequel effectuer l'analyse. Ce choix doit être motivé par la présence d'un nombre suffisant de conduites et de défaillances sur chaque classe de conduite (la classe étant l'attribut que l'on souhaite oublier puis compléter, ici la DDP). Le secteur candidat le plus valide est donc Couillet, qui possède le plus de conduites (28 776) et le plus de défaillances de conduites (4 331) parmi les 17 secteurs disponibles. Des calages initiaux avec le logiciel Casses sur les 17 secteurs ont également révélé que Couillet et l'un des secteurs avec le plus grand nombre de covariables trouvées significatives et dont l'ajout ou non dans le modèle montre un impact visible sur sa performance, avec une qualité de calage (aire sous la courbe de validation, voir 4.3.1.2) parmi les plus élevées. Au final, Couillet a donc été conservé comme secteur de référence, et les périodes suivantes ont été utilisées :

- Période de calage : 03/01/2011 – 31/12/2018.
- Période de validation : 01/01/2019 – 31/12/2021.

Le calage de référence est résumé dans le Tableau 29. La section 4 fournit des explications sur le processus de calage et de validation avec le LEYP.

Tableau 29 : Calage du modèle LEYP de référence sur Couillet.

Coefficient	Phénomène/caractéristique pris(e) en compte	Valeur	p-value	Commentaire
α	défaillances passées du tronçon	4.73	0	
δ	âge du tronçon	1	NA	Valeur fixée car non significative
β_{DIA}	diamètre	-0.0018	0	
β_{DDP}	date de pose (année)	0.010	0	
$\beta_{\ln(LNG)}$	longueur du tronçon (en logarithme naturel)	0.545	0	
β_{FG+inc}	matériau de référence = fonte grise + inconnus	0	NA	Valeur fixée car référence
β_{FD}	matériau fonte ductile	-0.269	0	
β_{PVC}	matériau PVC	-0.2291	0	
β_{AMCI}	matériau amiante ciment	-0.2307	0	
β_{PE}	matériau polyéthylène	0.2264	0	

β_{ACIER}	matériau acier	- 0.08628	0.037	
------------------------	----------------	--------------	-------	--

Attribut sélectionné pour la complétion

Le calage du modèle LEYP avec Casses a montré que l'âge du tronçon n'est jamais une variable significative (paramètre δ fixé à la valeur 1) pour l'ensemble des secteurs, y compris Couillet. Une hypothèse est que l'incertitude de la DDP dans les données d'origine (estimation à partir des matériaux qui sont eux-mêmes déduits des diamètres, et arrondi à la décennie) pourrait partiellement masquer l'effet du vieillissement des tronçons sur le calage. Cette incertitude est bien sûr à ajouter aux autres, telles que l'affectation des défaillances historiques aux bons tronçons du SIG. De plus, le calage a révélé qu'il est difficile d'apprécier l'effet de la complétion de la DDP sur le modèle, lorsque seuls des indicateurs de « validation » du modèle sont utilisés, tels que l'aire sous la courbe (voir 4.3.1.2 pour la description de la validation du modèle LEYP). Afin de travailler avec un attribut des conduites significatif pour le LEYP (p -value < 0.05), et dont l'omission ou l'inclusion dans le modèle possède un effet visible sur le calage (d'où un impact potentiel de sa complétion ou non par diverses méthodes), nous avons choisi de définir des scénarios d'oubli et de complétion de données se focalisant sur le matériau au lieu de la DDP.

Scénarios employés

Les scénarios suivants se basent sur l'omission de 40 % des matériaux sur le réseau de Couillet. Ce seuil de 40 % a été choisi pour correspondre à une valeur suffisamment élevée et réaliste, ici proche du taux de DDP manquantes sur le réseau français (Husson et al., 2018).

- Scénario de référence : toutes les données sont connues à 100 % (MAT, DIA, DDP).
- Scénarios d'oublis :
 - 40 % de matériaux oubliés et non complétés.
 - 40 % de matériaux oubliés et complétés :
 - par la méthode c1 ;
 - par la méthode c2 en utilisant le diamètre comme critère de propagation ;
 - par la méthode m1 (mode) en utilisant le diamètre pour faire des populations de conduites à partir desquelles les modes sont calculés.

Deux méthodes d'oubli des matériaux ont été employées, une totalement aléatoire, et une par ICAR de rue. Au total, il y a donc 9 scénarios et 9 calages du LEYP associés. Enfin, de la même façon que pour la validation des méthodes de complétion, afin de prendre en compte l'effet aléatoire de l'oubli de données et éviter les conclusions dues au hasard, pour chacun des 9 scénarios ci-dessus sauf pour celui de référence, 100 essais d'oubli de 40 % des matériaux ont été effectués, suivis des complétions et des calages associés. Il s'agit donc d'une simulation de Monte Carlo.

Indicateurs de comparaison des calages

L'impact de la complétion de données sur les prédictions du LEYP a été mesuré par 4 critères qui comparent le scénario de référence et les autres scénarios. Dans tous les cas, les scénarios autres que celui de référence sont associés à 100 essais d'oublis de matériaux et de complétion (Monte Carlo).

L'aire sous la courbe a été employée comme le critère classique concernant la performance au sens de la validation du modèle. À cela a été ajouté le coefficient α du LEYP qui mesure le poids des défaillances

passées sur l'intensité de défaillance d'un tronçon. Ainsi, un α plus élevé traduit généralement une capacité réduite du modèle à expliquer les défaillances par les autres covariables du tronçon, et donc un pouvoir prédictif plus faible.

Concernant l'écart des prédictions au calage de référence, les valeurs des $n=11$ paramètres θ_k du LEYP ont été comparées aux valeurs de référence $\theta_{ref,k}$ (voir Tableau 29) selon la formule suivante pour chaque scénario :

$$\Delta_{par} = \frac{1}{n} \sum_{k=1}^n 100 \times \left| \frac{\theta_k - \theta_{ref,k}}{\theta_{ref,k}} \right|$$

De plus, les rangs des tronçons classés selon le risque prédit de défaillance ont été comparés entre la référence (r_{ref}) et chaque scénario (r) pour l'ensemble des 28 776 tronçons et des 100 essais d'oublis. Ceci a été fait en inspectant directement la distribution de l'écart en rang, calculé selon la formule suivante pour chaque tronçon :

$$\Delta_{rang} = 100 \times \left| \frac{r - r_{ref}}{r_{ref}} \right|$$

Comparaison des calages entre scénarios

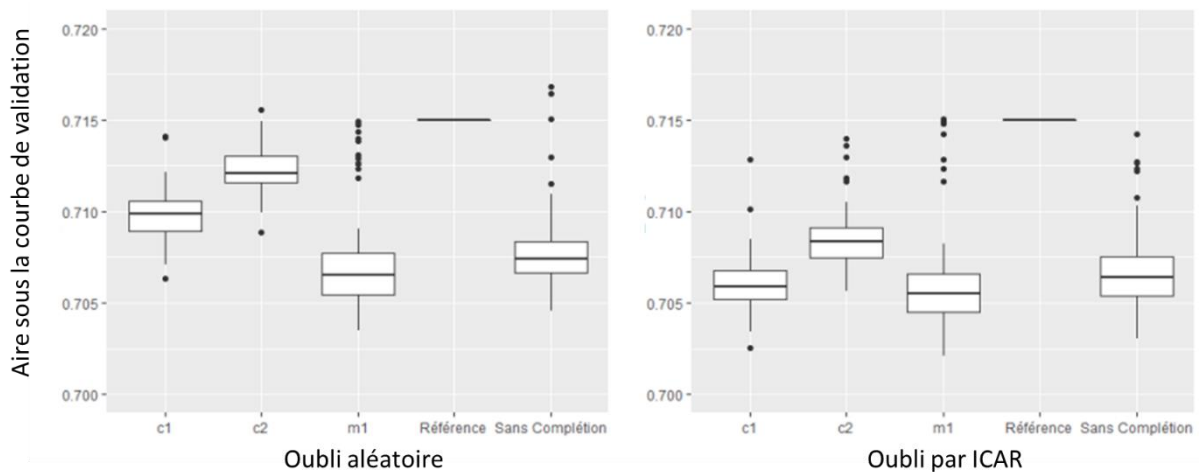


Figure 35 : Impact de l'oubli et de la complétion de 40 % des matériaux sur l'aire sous la courbe de validation du LEYP.

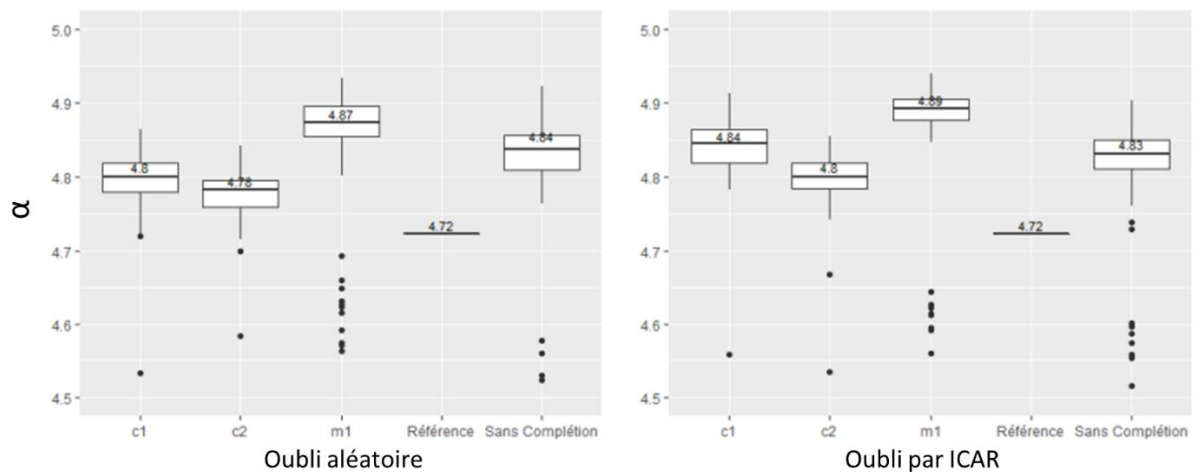


Figure 36 : Impact de l'oubli et de la complétion de 40 % des matériaux sur le coefficient α du LEYP.

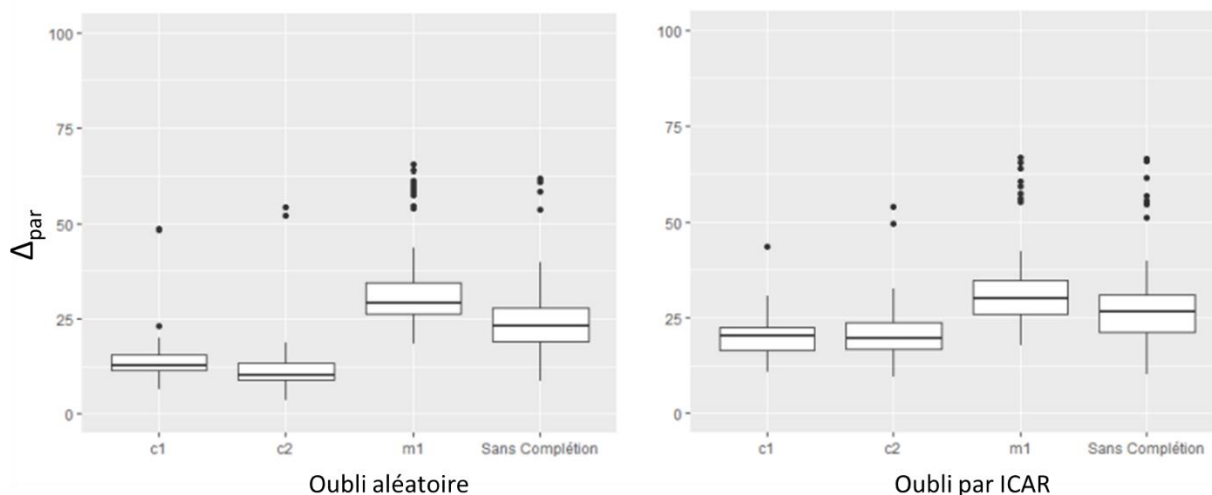


Figure 37 : Impact de l'oubli et de la complétion de 40 % des matériaux sur les paramètres du LEYP.

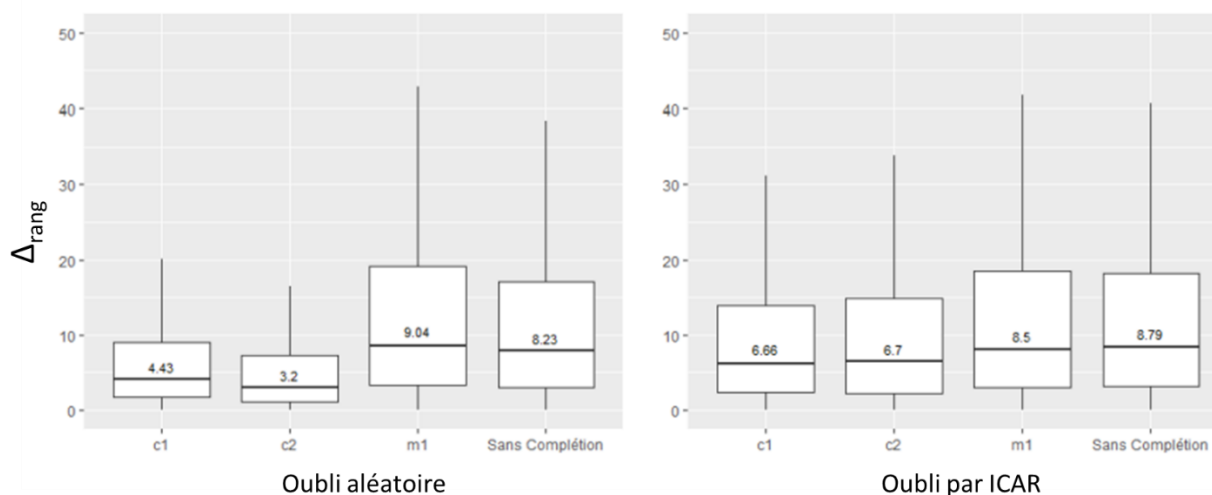


Figure 38 : Impact de l'oubli et de la complétion de 40 % des matériaux sur les rangs des tronçons classés par risque de défaillance par le LEYP.

L'hypothèse formulée en début de chapitre a donc été confirmée (Figure 35 et Figure 36) : une information déjà bien connue à l'origine (scénario de référence) est supérieure à une information bien complétée (méthodes c1 et c2), cette dernière étant supérieure à une absence d'informations (oubli sans complétion), cette dernière étant supérieure à une information « mal complétée » (méthode m1). De plus, une complétion de meilleure qualité permet de se rapprocher des paramètres du LEYP qui seraient obtenus avec les données d'origine (Figure 37 et Figure 38) et donc des prédictions des défaillances comme si tous les attributs des conduites étaient connus à l'origine. La méthode d'oubli des données a cependant une influence non négligeable sur l'impact des méthodes de complétion sur le calage du LEYP.

Conclusion

De façon générale pour un réseau d'eau potable, la meilleure solution pour compléter les attributs manquants de certaines conduites semble être de combiner la méthode qui utilise l'adjacence des conduites (méthode c1), avec des règles de cohérence entre les caractéristiques de conduites

(notamment entre DDP et MAT et entre MAT et DIA, ce qui donne la méthode c2). Cette approche combinée permet de se rapprocher des données d'origine et du calage du LEYP avec ces dernières. La méthode plus élaborée qui effectue une complétion de proche en proche (de type c1) par modalité d'un attribut donné (méthode c3) peut être envisagée à la place de la méthode c2 si un temps de calcul important n'est pas un frein.

Annexe 7. Note méthodologique sur les modèles LEYP et LEYP2s

1 Le modèle LEYP de défaillances répétées

Comme exposé par Le Gat [2014], les travaux internationaux consacrés au phénomène de défaillance des conduites d'eau sous pression s'accordent sur les points suivants :

- le taux instantané de défaillance (nombre de défaillances susceptibles d'affecter une canalisation par unité de temps) tend à augmenter avec l'âge de la canalisation ;
- à un âge donné, le taux de défaillance varie selon des caractéristiques propres à la canalisation (longueur, diamètre, matériau, type de joints, *etc.*), ainsi que selon la nature de son environnement (position sous chaussée ou sous trottoir, trafic dans la voie, nature du sol encaissant, *etc.*) ou de son exploitation (pression de service, variation de cette dernière, agressivité de l'eau transportée, *etc.*) ;
- les défaillances présentent une certaine agrégation spatiale, qui se traduit par leur tendance à s'accumuler sur les mêmes canalisations, le délai entre deux défaillances successives tendant en outre à se raccourcir quand le rang de la défaillance augmente.

Cela a conduit à proposer un modèle probabiliste du processus de défaillance issu du cadre théorique des *processus stochastiques de comptage* (cf. Ross [1996]), basé sur le processus de Yule (cf. Greenwood and Yule [1920]), et conçu comme une synthèse des approches précédemment proposées par Eisenbeis [1994], Pelletier [1999] et Røstum [2000].

Comme illustré par la figure 1.1, le processus de défaillance d'une canalisation, dont l'âge est noté t , est formalisé par la fonction de comptage $N(t) \in \mathbb{N}$, qui prend la valeur 0 à $t = 0$ (*i.e.* à l'installation de la canalisation), et s'incrémente d'une unité à chaque défaillance ; la différentielle $dN(t)$ de $N(t)$ est nulle partout sauf à chaque instant de défaillance t_j , où elle vaut 1.

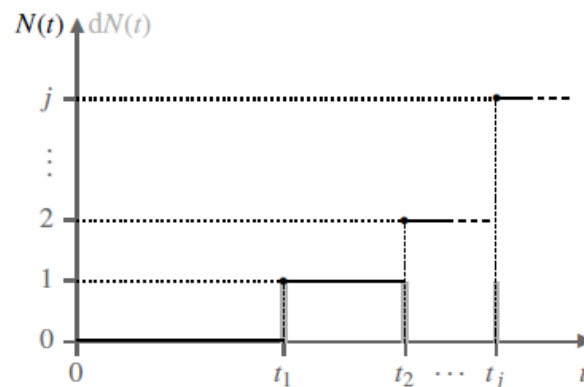


Figure 1.1: Fonction de comptage $N(t)$ et sa dérivée $dN(t)$

L'Extension Linéaire du Processus de Yule, *Linear Extension of the Yule Process* (LEYP), est définie par son intensité, qui est l'espérance conditionnelle de la différentielle de $N(t)$:

$$\forall t \in \mathbb{R}_+, \alpha \in \mathbb{R}_+^* : \\ E(dN(t) | N(t-), Z) = (1 + \alpha N(t-))\lambda(t, Z)dt$$

où $N(t-)$ est le nombre de défaillances subies par la canalisation entre sa pose et l'instant précédant juste l'âge t , et $\lambda(t, Z)$ est une fonction réelle de t et des caractéristiques de la canalisation rassemblées dans le vecteur Z (covariables).

Le LEYP contient comme cas particuliers le processus de Poisson non homogène (NHPP), pour $\alpha = 0$. A noter que le processus de Poisson a la propriété importante d'être sans mémoire de son passé, contrairement au processus de Yule ($\alpha > 0$).

Alors que la fonction de comptage du Processus de Poisson suit une distribution de Poisson, celle du LEYP suit une loi binomiale négative :

$$\forall t \in \mathbb{R}_+, N(t) \sim \mathcal{NB}(\alpha^{-1}, \mu(t)^{-1})$$

où, pour alléger la notation :

$$\mu(t) = \exp\left(\alpha \int_0^t \lambda(u) du\right)$$

La distribution binomiale négative est conservée pour le nombre de défaillances susceptibles de se produire dans un intervalle de temps $[c, d]$, connaissant le nombre de défaillances qui se sont produites dans un intervalle antérieur $[a, b]$, comme illustré par la figure 1.2 :

$$[N(d) - N(c) \mid N(b) - N(a) = m] \sim \mathcal{NB}\left(\alpha^{-1} + m, \frac{\mu(b) - \mu(a) + 1}{\mu(d) - \mu(c) + \mu(b) - \mu(a) + 1}\right)$$

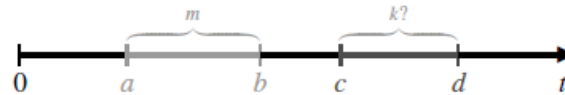


Figure 1.2: Fenêtres d'observation et de prédiction

Cette propriété est importante pour l'utilisation pratique du modèle LEYP, car elle permet de prédire explicitement le nombre de défaillances futures en se basant sur une connaissance possible incomplète de l'historique de défaillances (*i.e.* quand $a > 0$), ce qui le plus souvent le cas pour des canalisations anciennes, seulement observées sur une fenêtre temporelle récente et étroite. Cela permet en particulier de classer les tronçons d'un réseau par espérances* décroissantes des nombres de défaillances possibles, aux fins de simulation de rénovations ciblées.

*L'espérance d'une variable aléatoire de distribution binomiale négative $\mathcal{NB}(\theta, p)$ est calculée comme $\frac{\theta(1-p)}{p}$

2 Le modèle conjoint LEYP2s de défaillances répétées et de mise hors service

Le modèle LEYP exposé ci-avant en section 1 suppose implicitement que les objets dont le processus de défaillance est considéré ont une durée de vie infinie. Cette hypothèse a un double intérêt :

- théorique, de simplification du phénomène, aux fins d'étude de la distribution du processus de défaillance,
- pratique, de mise en œuvre de la modélisation en conditions d'information dégradée, lorsque la nature des données ne permet pas de considérer la durée de vie des objets étudiés (comme dans le cas des branchements).

Cependant ignorer la possible interdépendance entre le processus de défaillance et celui de mise hors service (MHS) présente le risque de biaiser le calage du modèle de défaillance, du fait de la non prise en compte du phénomène de *survie sélective*. Le biais de survie sélective découle directement de :

- la dépendance de l'intensité de défaillance du nombre de défaillances passées,
- la dépendance partielle de la MHS des défaillances répétées (les rénovations dites ciblées et d'opportunité),
- l'incomplétude d'une majorité des chroniques de défaillances disponibles.

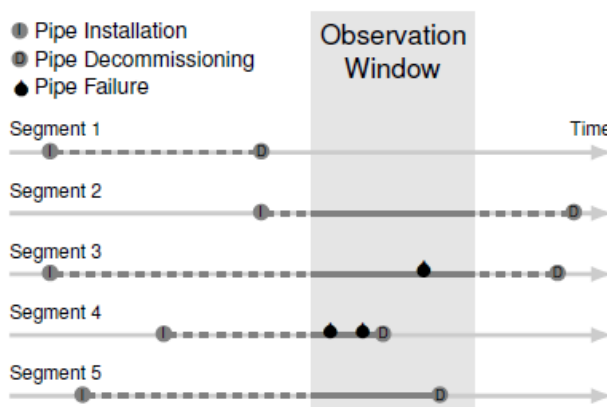


Figure 2.1: Historique incomplet et biais d'information

Les chroniques de défaillances observées concernent ainsi le plus souvent une population incomplète de canalisations, sélectionnées pour partie sur leur robustesse, comme illustré par les tronçons 1, mis hors service avant de pouvoir être observé, et 3 de la figure 2.1, posés la même année. De même, l'observation du tronçon 4 est interrompue prématurément par sa MHS pour cause de défaillances répétées, comparé aux tronçons 3 et 5.

La figure 2.1 illustre ainsi les deux biais d'information auxquels sont susceptibles d'être exposés nombre de chroniques de défaillances observées de canalisations :

- la troncature à gauche, lorsque des cohortes de canalisations sont incomplètement observées,
- la censure à droite de l'observation du processus de défaillance par la survenue de la MHS.

Le biais de calage du modèle de défaillance provient du caractère *informatif* de la troncature et de la censure, qui dépendent au moins partiellement du phénomène modélisé ; le modèle ainsi biaisé aura tendance à sous-estimer la probabilité de défaillance.

Afin de pallier ces biais d'information, Le Gat [2016] propose la modélisation conjointe des processus de défaillance et de MHS. Le processus de MHS est formalisé pour ce faire par la fonction de comptage $R(t)$ dont l'intensité dépend partiellement de $N(t-)$. Le couplage résultant est nommé *Linear Extension of the Yule Process with selective survival* (LEYP2s) :

$$\forall t \in \mathbb{R}_+, \begin{cases} N(0) = 0 \\ E(dN(t) | N(t-)) = (1 + \alpha N(t-)) \lambda(t) dt \\ R(0) = 0 \\ E(dR(t) | N(t-)) = (\psi(t) + \phi N(t-)) dt \end{cases}$$

with: $\alpha > 0$, $\lambda(t) \geq 0$, $\psi(t) \geq 0$, $\phi > 0$

L'intensité de MHS dépend pour partie d'une fonction $\psi(t)$ de l'âge de la canalisation, et possiblement de covariables (en particulier la longueur du tronçon), et pour partie de la dégradation de la canalisation, au travers du paramètre ϕ .

Le calage du modèle LEYP2s nécessite de disposer d'une chroniques de défaillances et mises hors service observées sur la même fenêtre, et repose sur une propriété importante de ce processus théorique, relative à la distribution conditionnelle du nombre de défaillances susceptibles de se produire dans un intervalle de temps futur, connaissant le nombre de défaillances intervenues dans la fenêtre d'observation, et sachant que la canalisation était encore en service au début de cette fenêtre. Comme pour le LEYP, cette distribution est binomiale négative :

$$[N(d) - N(c) | N(b) - N(a) = m, R(a-) = 0] \sim \mathcal{NB}\left(\alpha^{-1} + m, \frac{\mu(b) - \nu(a)}{\mu(d) - \mu(c) + \mu(b) - \nu(a)}\right)$$

$$\text{où } \nu(a) = \int_0^a e^{-(a-t)\phi} d\mu(t)$$

Comme dans le cas du LEYP, cette propriété du LEYP2s permet la prédiction des défaillances futures, en exploitant toute l'information disponible, même limitée à une fenêtre d'observation étroite, tronquée et censurée par le processus de défaillance lui-même.

Il est constaté empiriquement que l'absence de correction du biais de survie sélective n'empêche pas le modèle LEYP de produire des prédictions à court terme du taux de défaillance qui sont suffisamment précises pour classer de façon pertinente les conduites en vue de programmer les renouvellements annuels. La non-correction de ce biais devient cependant gênante dans le cadre de simulations à moyen ou long terme, car les estimations des paramètres α et δ portent alors implicitement l'effet des règles passées de gestion du patrimoine, alors que l'objet même des simulations long terme est d'évaluer l'effet de règles différentes ; l'utilisation du modèle LEYP2s prend alors tout son sens.

Annexe 8. Regroupement des prestations de RF en campagnes

Principe du regroupement

Le regroupement en campagnes employé exploite le fait que des prestations proches dans le temps et se déroulant sur une même cellule font vraisemblablement partie de la même campagne, en l'absence d'autres informations sur la nature exacte des prestations. Le regroupement est donc fondé sur le calcul des délais entre prestations et de leur comparaison à un seuil donné.

Il a été choisi d'arrondir les délais considérés entre prestations à la journée. De plus, les prestations de RF « nettoyées » qui ont été à cheval sur plus d'une journée ont été scindées en prestations journalières. Puis, les prestations qui se sont déroulées en parallèle sur une cellule (mêmes dates mais par des agents différents) ont été fusionnées, en ajoutant les heures prestées (près de 30 000 fusions ainsi effectuées). Enfin, le délai en jours ouvrés entre chaque prestation et la suivante a été calculé (exemple : une prestation finissant le lundi soir est distante de 1 jour ouvré de la prestation le mardi matin). Ce calcul des durées en jours ouvrés exclut donc les samedi et dimanche, et les jours fériés.

Toutes les prestations successives dont les durées inter-prestation calculées sont strictement inférieures à un seuil s sont assignées à la même campagne (en pratique elles reçoivent le même identifiant de campagne NuméroCellule_NuméroCampagne). L'indicateur construit pour fixer la valeur du seuil s à une valeur optimale est le suivant :

$$I_c(s) = \frac{n}{m(s) \sum_j \delta_j(s)}$$

Avec :

- n : nombre total de prestations
- s : durée seuil (jours ouvrés) entre prestations distinguant deux campagnes distinctes
- $m(s)$: nombre de campagnes distinguées par s
- j : numéro de campagne
- δ_j : nombre de jours ouvrés de la campagne j

Choisir la valeur optimale de s revient à maximiser I_c qui représente un compromis entre le nombre de campagnes distinguées et la « densité » des campagnes (nombre de prestations/nombre de jours de campagne). L'avantage d'un tel indicateur est la forme concave de sa courbe (Figure 39). Par exemple, regrouper toutes les prestations en 1 campagne minimise $m(s)$, ce qui fait monter I_c , mais cela rend la somme des jours ouvrés maximale, ce qui fait chuter plus fortement I_c . Au contraire, ne procéder à aucun regroupement minimise le nombre de jours ouvrés, mais maximise $m(s)$.

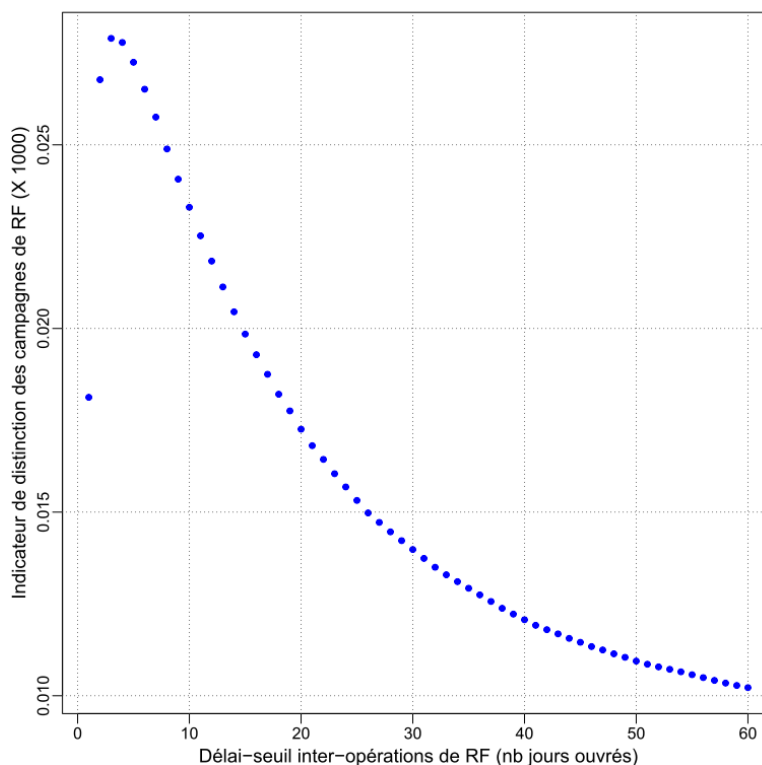


Figure 39 : Indicateur de délai-seuil inter-prestations optimal pour les données des 17 secteurs.

Pour l'ensemble des 17 secteurs, la valeur optimale est $s=3$ jours ouvrés (Figure 39). Ce délai peut être considéré court du point de vue opérationnel. Par exemple, on peut se demander si une prestation un lundi et une prestation le jeudi de la même semaine ne font pas finalement partie de la même campagne. Néanmoins des valeurs faibles de délai-seuil sont présentes sur la majorité des secteurs pris individuellement (Tableau 30). Il existe aussi des prestations une à deux semaines avant les campagnes qui en réalité correspondent probablement à du travail préparatoire de la campagne (comme la pose de prélocalisateurs acoustiques ou le repérage du réseau). Les fichiers de RF de la SWDE ne permettent cependant pas de distinguer systématiquement ces prestations des autres pour les regrouper avec leur campagne respective.

Tableau 30 : Délai-seuil optimal selon le secteur

Secteur	Délai-seuil optimal en jours ouvrés
5210	5
5220	4
5110	4
5440	3
5420	3
5810	3

5120	6
5230	5
5310	3
5330	4
5720	3
5510	3
5530	3
5610	3
5710	3
5820	3
Tous combinés	3

Enfin, il est clair qu'une partie des prestations de RF correspond à de la localisation de crashes par l'équipe RF (après passage de la première ligne), qu'il faudrait en théorie exclure de la notion de campagne systématique de RF (pilotée par la dérive des débits et non la signalisation par un tiers). Cependant le résultat des prestations de RF n'est pas précisé dans les données, et il doit être calculé, ce qui dépend donc de la définition des campagnes.

Autres indicateurs possibles

D'autres indicateurs ont été testés pour définir la valeur seuil optimale avant le choix de l'indicateur ci-dessus :

- une analyse graphique de l'histogramme des durées inter-prestations sur les secteurs de MAG ;
- une analyse de survie (Kaplan-Meyer) des durées inter-prestations sur le secteur de Mons ;
- une analyse graphique de la réduction du nombre de campagnes selon le délai-seuil, par classes de cellules ayant un nombre similaire de campagnes par durée d'observation totale.

Tous les indicateurs ci-dessus ont révélé qu'un délai-seuil de quelques jours à une semaine est optimal (selon les secteurs considérés et la méthode). En conclusion, le choix du délai-seuil, en l'absence de données plus opérationnelles pour la définition des campagnes, peut faire appel à de nombreuses méthodes, mais donne des résultats cohérents dans l'ensemble, bien que pouvant être considérés comme faibles.

Annexe 9. Mise à jour du calcul des pressions

Une pression de service élevée constitue un facteur pouvant altérer la durée de vie des canalisations. En première approche, les cotes MNT de chaque conduite de la cellule ont été calculées à partir du Modèle Numérique de Terrain ERRUISSOL (MNT à la maille de 10 mètres du Service Public de Wallonie) pour évaluer la dénivelée entre la conduite et le point haut de chaque cellule.

Ces données ont ensuite été combinées avec les mesures *in situ* sur les hydrants et les vannes de régulation pour calculer des indicateurs de pression lorsque c'était possible. Les cotes MNT du Trop Plein (TP) des réservoirs (extraites du fichier SWDE : OUVRAGE_CELLULE.xlsx) ont été exploitées en complément pour affiner la HMT (Hauteur Manométrique Totale) estimée de la cellule.

La couche de référence est fournie par le Service Public de Wallonie (SPW) à la maille 10 mètres (ERRUISSOL). Les altitudes ont été affectées à chaque extrémité des tronçons en croisant le MNT et la couche SIG des tronçons de canalisations agrégés. Pour approcher la pression exercée sur chaque tronçon, les cotes des nœuds d'extrémité, leur moyenne ainsi que les cotes minimum et maximum de la cellule d'appartenance du tronçon ont été ajoutées aux attributs des tronçons :

$$MNT_T = (MNT_1 + MNT_2)/2$$

Avec :

- MNT_T cote au sol attribuée au tronçon
- MNT_1 cote au sol du premier nœud d'extrémité du tronçon déduite du MNT du SPW
- MNT_2 cote au sol du second nœud d'extrémité du tronçon déduite du MNT du SPW

Calcul de la pression en présence d'hydrants

Pour le calcul de la pression, en l'absence de vannes de régulation, la moyenne des mesures disponibles réalisées sur les hydrants d'une cellule a permis d'estimer la HMT de celle-ci puis d'en déduire un indicateur de la pression exercée sur chaque tronçon par différence avec la cote au sol du tronçon telle que calculée précédemment.

$$HMT_C = \frac{\sum_{i=1}^n (MNT_{HYi} + P_{HYi})}{n}$$

Les conditions suivantes ont alors été appliquées :

- Si $HMT_C \leq MNT_{Cmax} + 5 m$ alors $HMT_C = MNT_{Cmax} + 5 m$
- Si des réservoirs sont présents sur la cellule, la cote MNT du trop plein du plus haut réservoir est noté MNT_{TP} . Une correction à HMT_C a été apportée dans les cas suivants :
 - 1^{er} cas : $HMT_C < MNT_{TP} \leq HMT_C + 5 m$ alors $HMT_C = MNT_{TP}$
 - 2^{ième} cas : $MNT_{TP} > HMT_C + 5 m$ alors la HMT_C a été augmentée de 5 m

Il vient ensuite le calcul de la pression de chaque tronçon :

$$P_T = HMT_C - MNT_T$$

Avec :

- HMT_C hauteur manométrique totale moyenne de la cellule
- MNT_{HYi} cote au sol de l'hydrant i de la cellule

- P_{HYi} pression mesurée pour l'hydrant i de la cellule
- n nombre d'hydrants de la cellule
- MNT_{TP} cote trop plein du réservoir le plus haut de la cellule
- MNT_{Cmax} cote au sol du nœud le plus élevé de la cellule
- P_T pression estimée du tronçon T

Calcul de la pression en présence de régulation, avec ou sans hydrants

Lorsqu'une vanne de régulation de cote au sol MNT_{VR} se trouve dans la cellule, toutes les mesures de pression des hydrants dont la cote moyenne est supérieure à celle de la vanne ($MNT_{HY} > MNT_{VR}$) ainsi que la pression amont de la vanne ont été utilisées pour calculer la valeur moyenne de la HMT à l'amont de la vanne (HMT_{cam}). Le cas échéant, les réservoirs ont été utilisés comme précédemment pour affiner HMT_{cam} pour l'étage le plus haut. Alors, pour tous les tronçons de canalisations dont la cote moyenne est supérieure à celle de la vanne de régulation ($MNT_T > MNT_{VR}$) la pression a été estimée par :

$$P_T = HMT_{cam} - MNT_T$$

Selon le même principe, la HMT à l'aval de la vanne de régulation (HMT_{cav}) a été calculée avec les hydrants dont la cote moyenne est inférieure ou égale à celle de la vanne ($MNT_{HY} \leq MNT_{VR}$). Pour tous les tronçons de canalisation dont la cote moyenne est inférieure ou égale à celle de la vanne de régulation ($HMT_T \leq HMT_{VR}$) la pression a été estimée par :

$$P_T = HMT_{cav} - MNT_T$$

Lorsqu'il y a plusieurs vannes de régulation dans la cellule, elles ont été considérées en cascade, les étages de HMT ont été déterminés par les cotes de chaque vanne de régulation. Pour chaque étage, toutes les mesures de pression des hydrants dont la cote moyenne est comprise entre la cote de deux vannes de régulation ainsi que la pression amont de la vanne la plus basse ont été utilisées pour calculer la valeur moyenne de la HMT de l'étage. La pression a alors pu être calculée pour tous les tronçons de canalisation dont la cote moyenne se situe entre les deux vannes de régulation.

A l'échelle de la cellule, une pression moyenne a été calculée à partir de la moyenne des pressions des tronçons, pondérée par la longueur des tronçons :

$$P_{moyC} = \frac{\sum_{i=1}^n P_{Ti} \times L_{Ti}}{\sum_{i=1}^n L_{Ti}}$$

Avec :

- P_{moyC} pression moyenne de la cellule
- P_{Ti} pression estimée du tronçon T_i
- L_{Ti} longueur du tronçon T_i
- n nombre de tronçons de la cellule

Calcul de la pression en l'absence de régulation et d'hydrant

Si aucune mesure de pression n'est disponible pour une cellule, sa HMT a été calculée en ajoutant 20 mètres colonne d'eau (mce) (valeur forfaitaire usuelle de la pression minimum admise qui peut être adaptée) à la cote au sol du nœud le plus élevé de la cellule.

$$HMT_C = MNT_{Cmax} + P_{Cminref}$$

En présence d'un réservoir, la formule de calcul a pu être modifiée, si $MNT_{TP} \geq MNT_{Cmax} + 5 m$, alors :

$$HMT_C = MNT_{TP}$$

- HMT_C hauteur manométrique totale de la cellule
- MNT_{Cmax} cote au sol du nœud le plus élevé de la cellule
- $P_{Cminref}$ pression minimum de référence de la cellule (par défaut 20 mce)
- MNT_{TP} cote trop plein du réservoir le plus haut de la cellule

Les pressions des tronçons et la pression moyenne de la cellule ont alors été calculées comme précédemment.

$$P_T = HMT_C - MNT_T$$

Annexe 10. Notes méthodologiques de la tâche 06

Introduction

Les systèmes d'alimentation en eau potable (AEP) consistent en un ensemble d'ouvrages et installations destinés à transporter et distribuer l'eau potable. L'eau est prélevée dans le milieu naturel (eaux superficielles ou souterraines) puis transférée vers des usines de traitement pour produire une eau conforme aux normes de potabilité en vigueur. Elle est ensuite stockée dans des réservoirs qui permettent de satisfaire les pointes de consommation. Finalement, l'eau est acheminée via un réseau de canalisations enterrées sur lesquelles seront raccordés les branchements en vue d'une desserte des usagers à domicile. Néanmoins, seulement une partie du volume prélevé est effectivement utilisée en raison de l'existence des pertes d'eau, sur canalisations ou branchements, lors de la distribution.

Matérialisées sous la forme de fuites, ces pertes relèvent essentiellement d'un phénomène de dégradation des réseaux de distribution, et ont pour effet d'en baisser le rendement³. Plus précisément, elles proviennent soit des ruptures de canalisations (les casses ou « crash » selon la terminologie Wallonne), soit de la mauvaise étanchéité des conduites et accessoires. Dans le premier cas, les fuites sont visibles et entraînent une inondation plus au moins étendue en surface, ce qui permet d'aider à localiser l'endroit de la casse. Dans le second cas, les écoulements à travers les défauts d'étanchéité demeurent invisibles en surface durant un certain temps. Ces fuites, qui peuvent être détectables (via diverses méthodes dont la détection acoustique) ou non détectables (dites fuites diffuses), sont celles qui affectent le plus le rendement des réseaux en raison de leur persistance dans le temps⁴. Outre l'impact sur la qualité des réseaux, les fuites affectent la performance du service sur le plan environnemental puisqu'elles contribuent à accentuer les prélèvements sur la ressource en eau destinée à la consommation humaine. Dans un contexte de réchauffement climatique, d'urbanisation et d'imperméabilisation des sols, les fuites pourraient compromettre les capacités d'approvisionnement en eau potable. De fait, lutter contre ces pertes en réseau de distribution apparaît opportun au moins sur trois plans : (i) préserver quantitativement les ressources en eau ; (ii) réduire les consommations énergétiques en évitant de pomper, traiter et distribuer une eau qui n'arrivera jamais aux usagers ; (iii) aider à maîtriser les dépenses publiques.

La lutte contre les fuites s'appuie sur plusieurs actions qui peuvent être classées en deux catégories. La première correspond aux actions de maintenance consacrées en particulier à la recherche et réparation des fuites en vue de réduire leurs durées d'écoulement et les volumes d'eau perdus. Ces actions regroupent les interventions qui ont lieu lors de fuites visibles ou détectées, et les campagnes systématiques de recherche de fuites non visibles et détectables. La seconde catégorie relève de la gestion patrimoniale et porte sur les actions de réhabilitation/renouvellement des parties du réseau qui enregistrent les fréquences les plus élevées d'apparition des fuites, et ce afin de limiter le nombre de fuites sur le réseau. Ces différentes actions, en particulier le remplacement des canalisations existantes, requièrent des investissements conséquents qui pourraient entraîner une augmentation du prix de l'eau. Ainsi, le choix d'une action tient pour une grande part à son avantage pour le maintien d'un service « performant » pour un coût (économique et environnemental) acceptable socialement.

Dans ce contexte, le projet « Gestion Patrimoniale Multi-Echelles des réseaux d'eau potable » (GePaME) a pour objectif de développer des outils opérationnels pour l'aide à la hiérarchisation des

³ Cf. par exemple [ASTEE et AIF, 2012], [Renaud et *al.*, 2014].

⁴ Cf. par exemple [Chesneau, 2006].

actions de réduction des pertes en eau par fuites afin de rationaliser les dépenses qui en découlent. Il implique des chercheurs en sciences pour l'ingénieur et en sciences sociales appartenant à l'équipe GPIE (Gestion des Infrastructures liées à l'Eau), composante de l'unité de recherche ETBX (Environnement territoires et infrastructures) d'INRAE ; et un opérateur public, la Société Wallonne des Eaux (SWDE), principal fournisseur d'eau potable de la Wallonie en Belgique. La tâche 6 « Analyse des pratiques de recherche de fuites » constitue une contribution sociologique au projet. Elle vise à éclairer la démarche de recherche de fuites et son intégration dans la gestion quotidienne du réseau d'eau potable de la SWDE, à la fois comme outil d'exploitation du réseau de distribution et outil de gestion patrimoniale.

Cette note méthodologique s'organise en deux parties. La première partie revient sur les objectifs de la tâche 6 et des éléments de problématique. La seconde partie porte sur les méthodes d'enquête.

Rappel des objectifs de la tâche 6 et éléments de problématique

La tâche 6 « Analyse des pratiques de recherche de fuites » traite du milieu sociotechnique impliqué dans la démarche de recherche de fuites pour mieux comprendre les conditions de sa mise en œuvre. Cette démarche comprend des objets techniques (débitmètres, capteurs, corrélateurs acoustiques, amplificateurs, etc.), des préconisations relatives à l'utilisation de ces objets, des grandeurs à mesurer pour surveiller le réseau de distribution (débits, niveaux d'eau, pressions, volumes d'eau), des étapes à suivre (sectorisation, prélocalisation, localisation), des outils d'exploitation des données (par exemple, l'outil GED⁵), des bases de données (par exemple, SIG⁶ réseau, SAP⁷), et des organisations impliquées dans la mise en œuvre de cette démarche. Par milieu sociotechnique nous désignons l'environnement existant dans lequel la recherche de fuite est déployée⁸, à savoir :

- l'ensemble des objets et dispositifs techniques qui constituent le système d'alimentation en eau potable (canalisations de distribution, canalisations de branchement, réservoirs, équipements de métrologie, équipements de surveillance des réseaux, matériels de détection de fuites, modèles, bases de données, etc.) ;
- les organisations engagées dans l'émergence, la gestion et le maintien en l'état de ces dispositifs ;
- les normes réglementaires et techniques ainsi que les modèles économiques en vigueur ;
- les pratiques existantes.

De fait, le milieu sociotechnique est à la fois technique, organisationnel, social, économique et politique. La tâche 6 a pour objet de décrire et d'analyser le milieu sociotechnique associé aux dispositifs de recherche de fuites, c'est-à-dire le milieu sociotechnique nécessaire à l'existence de ces dispositifs et leur « concrétisation » [Simondon, e2002]. La concrétisation désigne ici le passage d'un dispositif expérimental associé à la vie de laboratoire à un dispositif banal ou « d'ordinaire » [Jouët,

⁵ Gestion Electronique de Documents.

⁶ Système d'Information Géographique.

⁷ Systems Applications and Products in Data Processing.

⁸ En ce sens, le milieu sociotechnique se rapproche du « contexte sociotechnique » défini par Madeline Akrich comme l'environnement où se rencontrent des réalités tout à fait hétérogènes (techniques, géographiques, physiques, économiques, politiques, etc.) et dans lequel l'objet technique est appelé à fonctionner [Akrich, 1998].

2000] inscrit dans le quotidien des gestionnaires et exploitants des réseaux d'eau potable⁹. Partant, quel est le milieu sociotechnique au sein duquel la recherche de fuite est insérée ? En quoi ce milieu permet-il sa concrétisation ? Quels sont les facteurs qui favorisent ou limitent ce processus ?

Cadre d'analyse mobilisé

Cette contribution à l'analyse du milieu sociotechnique dans lequel la recherche de fuites est mise en œuvre s'appuiera sur un cadre conceptuel qui envisage cette activité comme activité instrumentée ou « outillée » [Sigaut, 2012], c'est-à-dire mobilisant des objets et des dispositifs techniques. Les objets et dispositifs techniques correspondent aux éléments fabriqués par l'homme. Les dispositifs sont des assemblages d'objets. Le vocable « dispositif » permet de rendre compte de l'agencement d'entités hétérogènes qui structurent l'activité de recherche de fuites. Le qualificatif « technique » souligne la logique d'instrumentalité, c'est-à-dire la mise en œuvre de moyens en vue d'une fin [Peeters et Charlier, 1999].

Selon cette approche, il n'est d'activité ou d'action sans un ou plusieurs objets qui accroissent les dispositions corporelles et cognitives des individus. Dans le même temps, il n'y a pas d'objets en fonctionnement sans une ou plusieurs organisations constituées autour de leur conception, réalisation, maintenance et gestion, et qui sont légitimes pour constituer l'existence de ces objets en ressources (par exemple, le service de l'eau rendu aux usagers). Autrement dit, il ne peut y avoir de dispositifs techniques sans organisations et tout dispositif technique est aussi organisationnel [Toussaint, 2009 ; Toussaint et Vareilles, 2010]. A partir de ces définitions, notre cadre d'analyse peut se résumer en trois propositions :

- les objets techniques délivrent des licences d'action et constituent ainsi des offres en pratiques sociales ;
- les individus agissent « rationnellement » dans le sens où ils ont des raisons d'agir comme ils le font ;
- les objets sont des acteurs « non-humains » qui participent de l'action au même titre que les acteurs humains.

Les objets techniques, des offres en pratiques sociales

Les objets techniques, du fait de leur configuration, appellent plus au moins des pratiques démultipliant ou limitant les moyens d'action des individus qui les utilisent. Ils constituent ainsi des offres en pratiques sociales¹⁰. Par exemple, un bassin d'infiltration des eaux pluviales aménagé dans un parc, appelle à marcher, courir, jouer (seul ou en groupe) lorsqu'il est vide. Chaque nouveau dispositif technique renouvelle l'offre en pratiques sociales. Parmi ces pratiques, certaines sont visées par les concepteurs (au moment de l'élaboration du dispositif) ; d'autres débordent le « script » promu par la conception¹¹ [Akrich, 1987 ; 1990]. Les débordements vont du simple changement du spectre

⁹ Dans la description du processus de concrétisation, le premier dispositif est dit « abstrait » ou « artificiel », il existe en dehors du monde (son existence est chaotique) ; le second « concret », son mode d'existence se rapproche de celui d'un dispositif « naturel » ou spontanément produit [Simondon, e2002].

¹⁰ Sur le fondement de cette proposition, cf. par exemple : [Rabardel, 1985], [Akrich et al., 2006], [Toussaint, 2009], [Denouël et Granjon, 2011].

¹¹ Lorsqu'ils conçoivent des objets, les concepteurs définissent un ensemble de pratiques escomptables qui s'inscrivent dans un ou plusieurs scripts ou scénarios. Ces scripts s'appuient sur des hypothèses sur les milieux

des usages visés (le « déplacement »), jusqu'à l'utilisation de l'objet pour un dessein qui n'a rien à avoir avec ses usages prescrits (le « détournement ») [Akrich, 1998]. De cette manière, les dispositifs techniques constituent des « instruments » pour les individus dans le sens où ils permettent à celui qui en use de voir différemment le monde, de trouver les moyens d'y agir, et finalement de passer à l'action¹². Mais cette capacité de l'objet à être un instrument est variable. Elle dépend de son appropriation aux activités sociales. L'appropriation est envisagée ici selon deux axes.

Le premier renvoie à l'appréhension des objets par les individus et aux manières qu'ils ont de les utiliser. L'appropriation est ainsi corrélative d'apprentissages sociaux nécessaires au maniement des objets. En effet, l'utilisation d'un dispositif nécessite des tours-de-mains et par conséquent, l'acquisition de savoirs, de savoir-faire et des habiletés personnelles qui permettent de discerner l'offre contenue dans le dispositif (par la découverte de son contenu technologique et de son mode opératoire), mais aussi d'ouvrir vers de nouvelles possibilités d'action.

Le second axe tient compte de l'insertion des objets dans un ensemble sociotechnique plus large, comprenant d'autres objets en usage, des pratiques déjà formées, des régimes juridiques, économiques et politiques, des normes, et des organisations. L'insertion pérenne suppose des ajustements réciproques entre les objets techniques et le milieu existant (par exemple, évolutions des organisations, des normes, ou de l'objet en question ainsi que des moyens à mettre en œuvre pour favoriser ces évolutions). Ces changements sont nécessaires pour que le milieu prenne en charge pleinement les objets et que ceux-ci s'incorporent dans des pratiques sociales, se banalisent, et par là même, « se naturalisent » [Simondon, e2001].

Les individus, des acteurs « raisonnables »

Les individus engagés dans l'action sont raisonnables dans le sens où ils ont de « bonnes raisons » d'agir comme ils le font¹³. Ces raisons sont fondées en situation (dans des circonstances spatiales et temporelles et en interaction avec les autres et les objets présents) et s'attachent en général à différentes structures de références, des « grandeurs » [Boltanski et Thévenot, 1991], incorporées par les acteurs et à partir desquelles sont déterminées leurs actions : par exemple, les valeurs, les croyances, les connaissances, les habitudes, les convenances et les normes sociales/culturelles de comportement propres aux groupes sociaux et aux situations d'action¹⁴.

Dans le cas de la recherche de fuites, les manières de faire, renvoient à des façons de mobiliser les objets techniques et aux comportements « organisationnels » des acteurs, lesquels comportements sont particulièrement orientés par : les buts fixés à l'activité, la division du travail, les règles formelles et informelles de l'organisation, la hiérarchie entre les différents acteurs.

sociotechniques dans lesquels ces dispositifs seront intégrés et sur les situations d'action dans lesquelles ils seront mobilisés.

¹² « Les instruments rendent intelligible le monde et chaque instrument nouveau, en modifiant les modalités d'intelligibilité du monde, modifie sa signification et ouvre les portes à de nouvelles possibilités d'agir. Ils sont parties prenantes de l'activité cognitive de ceux qui en usent » [Toussaint, 2009, p. 491].

¹³ Cf. par exemple [Bourdieu, 2001] et [Boudon, 2003] sur l'acteur et l'action.

¹⁴ A chaque situation, l'individu ajuste son comportement sur celui des autres selon des convenances et les normes de comportement attendues de chacun pour tous.

Les objets, des acteurs à part entière

Dans leurs activités, les individus interagissent avec des individus ainsi que des objets et des dispositifs techniques. De cette manière, les objets participent de l'action, notamment comme instruments. A l'instar des individus, ils sont des acteurs « non-humains ».

Dans cette perspective, nous considérons les connaissances comme distribuées entre des acteurs humains et non-humains. Par exemple, dans le cas du travail de recherche de fuites, les savoirs sur lesquels repose cette activité sont distribués entre des individus (ingénieurs, chercheurs, techniciens, opérateurs chercheurs de fuites, etc.) et une multitude d'objets (bases de données, algorithmes, appareils de détection de fuites, modèles, rapports, plans, fiches d'instruction, manuels d'utilisation, etc.).

Ces acteurs non-humains sont des « objets intermédiaires » qui circulent d'un acteur à l'autre ou rassemblent autour d'eux plusieurs acteurs [Vinck, 1999 ; 2009]. Les objets intermédiaires revêtent une importance particulière pour l'action soit parce qu'ils la prolongent, soit parce qu'ils la transforment. Ils constituent des « traces » [Vermersch, e2019] et des « marqueurs » de l'activité [Vinck, 2009]. Ainsi, dans l'activité de recherche de fuites, plusieurs objets sont produits et mobilisés. Ils interagissent avec d'autres acteurs humains ainsi que non-humains. Au cours de leur passage d'une entité à l'autre, le contenu de ces objets se déforme. Ils contribuent ainsi à l'articulation progressive d'un « réseau de traduction » ou d'une « chaîne de traductions sociotechniques » associant différentes entités humaines et non-humaines [Latour et Woogler, e1988 ; Latour, e2005 ; Callon, 1989].

Partis pris de recherche et premières hypothèses

Ce cadre d'analyse conduit à relier la question de la banalisation du dispositif de recherche de fuites à son appropriation aux activités de gestion quotidienne du réseau d'eau potable. Cette appropriation tiendrait dans la capacité de ce dispositif à s'actualiser dans des usages existants ou à en créer de nouveaux (par exemple, mise au point de modèles de prédiction de fuites, construction de stratégies d'action), et ce faisant à trouver des organisations pour garantir son existence. Cette façon de poser le problème engage à porter une attention particulière aux objets : les dispositifs techniques impliqués dans la recherche de fuites et les objets intermédiaires qui circulent entre les acteurs. De fait, la tâche 6 considère la manière dont les activités de recherche de fuites s'effectuent et mobilisent un ensemble d'objets en prenant comme point d'entrée méthodologique les « traces » laissées par ces activités, principalement des « données ». Le suivi de ces données et leurs différentes traductions permet de restituer les activités liées à la recherche de fuites.

Les données sur la recherche de fuites sont ici considérées comme un enchaînement de traductions dont l'ensemble forme une chaîne de traductions à la fois techno-logique et socio-logique¹⁵ [Baati, 2021].

La chaîne de traduction techno-logique relie un ensemble de dispositifs techniques (ensemble technologique). Elle consiste à objectiver un état physique par des mesures (mesures de vibrations, de débits, de pressions, etc.). Ces informations sont générées par des programmes de traduction incorporés dans divers objets et dispositifs techniques (les savoirs « algorithmisés »). Elles apparaissent sur « les écrans » sous forme d'inscriptions ou traces qui peuvent être à vérifier, à transformer, à combiner avec

¹⁵ La traduction doit être entendue ici comme un processus de transformation.

d'autres traces, à rectifier, à archiver, etc. De ce fait, cette chaîne de traduction participe à projeter du sens sur les données produites.

La chaîne de traduction socio-logique implique des individus et des organisations, c'est-à-dire à la fois :

- les dispositifs organisationnels : organisation du travail de recherche de fuites, division, spécialisation des tâches, transmission interpersonnelle ;
- les conditions de travail : statut des agents, ergonomie des lieux de recueil, de transmission et d'interprétation des données, appropriation des instruments aux activités des agents ;
- les protocoles en vigueur et leur application ;
- les habitudes, les routines et les croyances propres aux organisations et aux individus (ou groupes d'individus).

L'ensemble de ces éléments permet de mettre en œuvre le système technique de recherche de fuites et d'interpréter les données produites afin de leur donner du sens. Cette attribution de sens participe de l'ancrage de la recherche de fuite dans son environnement sociotechnique et par là de sa banalisation.

Les échanges et réunions sur les « données RFU¹⁶ » nous ont permis de repérer certaines limites quant à la valorisation de ces données. Ces limites sont pour la plupart liées à la saisie des données, leur mise en forme, leur classement, leur enregistrement et leur archivage. Ces observations nous conduisent à formuler trois hypothèses¹⁷ sur les finalités et les modes d'appropriation de la démarche de recherche de fuites déployée :

1. Les limites observées résulteraient de décalages entre les capacités du dispositif technique de recherche de fuites et les finalités effectives de sa mise œuvre. Ces décalages sont susceptibles de se nicher à différents interstices du système de gestion et de recherche de fuites de la SWDE et peser sur la cohérence d'ensemble (objectifs gestionnaires variés voire contradictoires, appréhension différenciée du travail de recherche de fuites selon les agents et/ou les métiers, défauts ou « imperfections » de l'appareillage technique utilisé, appropriation des infrastructures préexistantes à la recherche de fuites, etc.).
2. La valorisation des données produites dépendrait des significations qu'elles revêtent pour les acteurs engagés dans leur production et exploitation, et de ce qu'ils projettent d'en faire. En l'état, il semblerait que la chaîne de traductions sociotechniques à l'œuvre dans la recherche de fuites permet à la SWDE de produire des données « valables » et « utilisables » pour servir les finalités administratives et de gestion courante qu'elle se fixe. Cependant, en l'état, ces données ne semblent pas suffisamment complètes et robustes pour être utilisées dans le cadre d'une gestion patrimoniale préventive du réseau potable de la SWDE, fondée sur une approche statistique de défaillance et de renouvellement des canalisations.
3. Ces limites révèlent dans une certaine mesure les ajustements nécessaires pour que le dispositif de recherche de fuite s'insère d'une manière stable et efficace dans le milieu sociotechnique

¹⁶ Recherche de Fuites.

¹⁷ Ces hypothèses restent à ce stade d'avancement de la tâche 6 encore un peu grossières et nécessitent d'être mieux informées.

existant, qu'il fonctionne à la hauteur de ses potentialités d'action et qu'il se pérennise et se naturalise au travers de pratiques stabilisées et routinières. Or, en l'état, cette « banalisation » apparaît inachevée en raison d'un déploiement incomplet du dispositif. En effet, la mise en œuvre de la recherche de fuites au sein de la SWDE s'est accompagnée de changements : constitution de compétences et de savoir-faire, évolution des pratiques de gestion quotidiennes du réseau, apparition de nouveaux métiers. Partant, le dispositif de recherche de fuites trouve à s'actualiser pour partie dans les pratiques de gestion quotidienne du réseau (pratiques techniques et administratives). Cette dynamique étant en cours, ce dispositif demeure en partie expérimental et ne peut éprouver pleinement de stabilisation et de fonctionnement optimal.

Méthode d'enquête

L'analyse des activités sociales soulève des difficultés méthodologiques [De Certeau, e1990]. Ces difficultés sont liées à la « quotidienneté » des pratiques observées [Begout, 2005] et au décalage fréquent entre ce que les acteurs font et savent faire, et ce qu'ils disent faire et savoir faire [Lahire, 1998]. Pour lever ce verrou, il convient de mettre en œuvre autant que possible des dispositifs d'enquête assurant les moyens d'une réactivation du « faire » : observations *in situ* et des entretiens auprès des acteurs de la SWDE.

Les verrous méthodologiques

Les acteurs sont sans cesse pris dans une multitude d'activités qui se chevauchent, se coupent s'entremêlent, etc. Ces activités s'inscrivent dans des cours d'action, qui font l'objet d'apprentissage et qui sont en partie « routinisés » ou allant de soi [Kerbarth-Orechioni, e1997]. Lorsqu'au cours d'un entretien l'acteur parle, il ne livre pas un matériau langagier déjà constitué, mais il le construit en parlant. En passant du « registre procédural (savoir-faire) » au « registre déclaratif (savoir-dire) » [Blanchet et Gotman, e2007], il met en forme et en ordre son expérience afin de la rendre explicite. Il opère ainsi une abstraction qui transforme son expérience en une sorte de synthèse ou généralisation de ses activités (« en général je fais ceci et cela... ») et peut en grande partie détruire ce qui fait la logique de ces activités, notamment de point de vue de leur enchaînement et leur succession effective dans le temps. De fait, à la question « qu'avez-vous fait... ? », les acteurs répondent rarement en détaillant leurs multiples (et, en définitive innombrables) activités. Ils se contentent ordinairement de mentionner les activités qu'ils perçoivent comme principales. Ces limites sont liées, d'une part, à la visibilité de leurs pratiques, et d'autre part, à leurs modes d'insertion dans le cours de l'action [Lahire, 1998, pp. 16-22].

En effet, engagé dans une action, un acteur n'est pas spontanément conscient de tout ce qu'il fait, de la manière dont il le fait, et de l'ensemble des gestes, savoirs et savoirs faire qu'il met en œuvre dans son activité. Ces pratiques et ces savoirs ne sont pas nécessairement objectivés dans le sens où ils ne sont pas habituellement catalogués et classés comme des pratiques ou des savoirs identifiables en tant que tels. Par conséquent, ils ne sont pas socialement clairement identifiés et nommés, ce qui les rend difficilement déclarables lors d'un entretien. Par exemple, si un enquêté est interrogé sur ce qu'il fait lorsqu'il recherche une fuite et qu'il répond « j'écoute le réseau », cela ne signifie pas que ce qu'il a réellement fait relève de l'écoute de vibrations. Il peut avoir fait un tour sur le terrain pour deviner le tracé des conduites, délimiter visuellement la zone à inspecter, géolocaliser le point de fuite trouvé, noter les caractéristiques de la fuite détectée, etc. Ainsi, lorsqu'un acteur raconte ce qu'il a fait, il a tendance à occulter certaines activités. De même, il occulte « les micro-pratiques » qu'il a effectuées et les « micro-savoirs » sur lesquels il s'est appuyé pour faire ce qu'il a fait [Lahire, 1998, p. 18]. Ces

micro-pratiques et micro-savoirs sont généralement considérés comme des moyens pour arriver à une fin représentée par une activité principale. Par exemple, lorsqu'un enquêté déclare avoir relié un capteur de niveau d'eau à une armoire électrique, il ne mentionne pas qu'il a lu un schéma électrique, un plan de câblage, qu'il a paramétré le capteur et donc lu un manuel de paramétrage, etc. C'est parce que les activités de lecture et de paramétrage ne sont dans ce cas perçues que comme des activités secondaires et annexes intervenant en support dans le cadre d'une activité principale.

La production par les acteurs de discours sur leurs pratiques dépend aussi de ce qui leur apparaît légitime par rapport à ce qu'ils imaginent que l'enquêteur attend [Beaud et Weber, e2003 ; Kauffman, e2004]. L'acteur interrogé peut tenir un discours « très positif » de ses pratiques, il n'évoque que rarement les problèmes qu'il rencontre [Diaz, 2005]. Il peut également formuler des réponses contraires à son comportement habituel et conformes à ce qu'il croit être l'attente de l'enquêteur. De fait, les comportements ne sont pas formulables n'importe où, ni dans n'importe quelle circonstance¹⁸ [Blanchet et Gotman, e2007, p.26]. Interviennent ainsi des jeux de rôles et de projections entre les personnes enquêtées et enquêtrices, qui agissent sur les discours et les récits [Bachir et Bussat, 2000 ; Cahborderon et al., 1994 ; Pinçon et Pinçon-Charlot, e2002].

Ces limites rendent malaisée l'appréhension des pratiques sociales (individuelles et collectives) par la production de données verbales. Elles invitent autant que faire se peut à mobiliser des méthodes permettant de saisir les activités sociales dans leur contexte « naturel » [De Sardan, 2001] et à les hybrider [Becker, 2002 ; Fabiani, 2003]. Dans notre situation, la reconstitution du déroulement des activités liées à la recherche de fuites incite à articuler : des observation *in situ* portant sur les pratiques mises en œuvre et les objets manipulés par les agents (en y participant ou non aux scènes observées¹⁹) ; et des entretiens afin d'inviter les acteurs à se remettre en situation d'action et à nous expliciter les logiques qui sous-entendent leurs pratiques.

Les dispositifs d'enquête

Ce travail s'appuie sur une enquête de terrain qui vise à renseigner les modalités de mise en œuvre de la recherche de fuites par la SWDE. L'enquête sera menée dans différentes « zones géographiques » de la SWDE. Cela ne vise pas à comparer le déploiement de la recherche de fuites dans les « zones » enquêtées, mais plutôt d'apporter des éclairages variés voire complémentaires sur les moyens et les manières d'agir dans une diversité de situations. Les personnes enquêtées (interviewées et observées) sont anonymisées. Les données recueillies seront tenues « confidentielles », au sens où elles resteront propriété d'INRAE et ne seront mobilisées que de manière anonyme.

Le protocole de recherche envisagé porte une attention particulière aux dispositifs techniques de recherche de fuites et aux objets intermédiaires qui circulent entre les acteurs engagés dans ces activités (OT, extraits de données, notes, etc.). Il comprend une recherche documentaire (par exemple, notes et documents techniques, rapports de stratégie et d'orientation, manuels, règlements intérieurs, rapports d'activités) ; des entretiens auprès des acteurs mobilisés dans la recherche de fuites ; et des observations de cette activité. A côté de ces dispositifs d'enquête formels, les réunions (réunions

¹⁸ Dit de manière triviale, on ne raconte pas n'importe quoi à n'importe qui n'importe où.

¹⁹ Par observation participante nous entendons un dispositif dans lequel l'observateur s'immerge dans le travail des enquêtés et partage leurs activités pour mieux les observer. L'observation directe des pratiques consiste à observer les comportements des individus dans les lieux de leurs activités sans en modifier le déroulement ordinaire. L'observateur est présent sur place, mais en retrait par rapport à l'action.

d'équipe et CSO) sont aussi l'occasion de recueillir des informations sur les pratiques de recherche de fuites de la part des praticiens et des chercheurs participants. Elles font l'objet de notes écrites dans un carnet de recherche.

Les entretiens auprès des acteurs de la recherche de fuites visent plus à faire décrire un état des choses (par exemple, décrire un processus de localisation d'une fuite, préciser un organigramme ou le rôle des individus et des dispositifs techniques mobilisés dans la recherche de fuites). Ils ont pour but de recueillir des informations factuelles sur l'organisation du travail au sein de la SWDE, sur le fonctionnement des dispositifs techniques, et sur les pratiques de recherche de fuites. Les entretiens seront enregistrés dans la mesure du possible à l'aide d'un dictaphone²⁰ et retranscrits de la manière la plus fidèle possible.

Les observations *in situ* traitent prioritairement des pratiques des trouveurs de fuites²¹, et mobiliseront le cas échéant la photographie²², les enregistrements audio, la prise de notes et/ou le croquis. Elles visent à appréhender l'activité « en train de se faire », en s'attachant notamment à :

- rendre compte des pratiques réellement en vigueur dans les activités de recherche de fuites ;
- décrire la place et le rôle des dispositifs techniques dans ces activités ;
- mieux connaître la manière dont sont mobilisés les dispositifs techniques de recherche de fuites dans les activités des acteurs (compétences cognitives et empiriques mises en œuvre dans la manipulation des objets, modes d'appropriation et intégration dans le quotidien des enquêtés).

Bibliographie de la note méthodologique

AKRICH Madeleine, « De la sociologie des techniques à une sociologie des usages », *Techniques et culture*, 1990, vol. 2, n° 3, pp. 83- 110.

AKRICH Madeleine, « La construction d'un système socio-technique. Esquisse pour une anthropologie des techniques », *Anthropologie et sociétés*, 1989, vol. 2, n° 13, pp. 31- 54.

AKRICH Madeleine, « Comment décrire les objets techniques? », *Techniques et culture*, 1987, n° 9, pp. 49- 46.

AKRICH Madeleine, CALLON Michel et LATOUR Bruno, *Sociologie de la traduction*, Presses Des Mines., Montreuil, Sciences Humaines, 2006, 304 p.

BAATI Selma, *Analyse des conditions de mise en oeuvre de l'autosurveillance des réseaux d'assainissement. Le cas des agglomérations de Lyon et de Nantes dans les années 1990 et 2010*, Lyon, Institut National des Sciences Appliquées de Lyon, 2021, 293 p.

BEAUD Stéphane et WEBER Florence, *Guide de l'enquête de terrain*, La Découverte., Paris, Guides Repères, 2003, 356 p.

BECKER Howard S., *Les ficelles du métier*, La Découverte., Paris, Guides Repères, 2002, 360 p.

BLANCHET Alain et GOTMAN Anne, *L'enquête et ses méthodes. L'entretien*, Paris, Collection 128, 2007,

²⁰ L'enregistrement fait toujours l'objet d'une demande au début de l'interview.

²¹ Les observations ne visent en aucune façon de « juger » de la qualité du travail réalisé, mais bien de comprendre les logiques propres aux outils utilisés, et les logiques des actions (protocoles, règles, connaissances empiriques, division des tâches, etc.).

²² Les visages dans les photos sont systématiquement floutés.

125 p.

CALLON Michel, *La science et ses réseaux. Genèse et circulation des faits scientifiques*, La Découverte., Paris, Anthropologie des sciences et des techniques, 1989, 214 p.

DE CERTEAU Michel, *L'invention du quotidien, I : Arts de faire*, Gallimard., Folio Essais, 1990, 416 p.

DE SARDAN Jean-Pierre Olivier, « L'enquête de terrain socioanthropique », in *Corpus, source et archives*, Boutier Jean, Fabiani Jean-Louis, De Sardan Jean-Pierre Olivier., Tunis, IRMC, 2001.

DENOÛËL Julie et GRANJON Fabien, *Communiquer à l'ère du numérique. Regards croisés sur la sociologie des usages*, Transvalor-Presses des Mines., Paris, Sciences Sociales, 2011, 320 p.

DIAZ Frédéric, « L'observation participante comme outil de compréhension du champ de ma sécurité. Récit d'un apprentissage de l'approche ethnographique pour tenter de rendre compte de la complexité du social », 2005, URL : <https://journals.openedition.org/champpenal/79>.

JOUËT Josiane, « Retour critique sur la sociologie des usages », *Réseaux*, 2000, vol. 18, n° 100, pp. 487- 521.

LATOUR Bruno, *La science en action. Introduction à la sociologie des sciences*, La Découverte., Sciences humaines et sociales, 2005, 672 p.

LATOUR Bruno et WOGLER Steve, *La vie de laboratoire. La production des faits scientifiques*, La Découverte., Sciences et Société, 1988, 296 p.

PINÇON Michel et PINÇON-CHARLOT Monique, *Voyage en grande bourgeoisie. Journal d'enquête*, Presses Universitaires de France., Quadrige essais débats, 2002.

SIMONDON Gilbert, *Du mode d'existence des objets techniques*, Aubier., France, Philosophie, 2001, 368 p.

TOUSSAINT Jean-Yves, « Les usages et les techniques », in *Traité sur la ville*, Stébé Jean-Marc, Marchal Hervé., Paris, Presses Universitaires de France, 2009, pp. 461- 512.

TOUSSAINT Jean-Yves et VAREILLES Sophie, « Handicap et reconquête de l'autonomie: réflexions autour du rapport entre convivialité des objets et autonomie des individus », *Geographie Helvetica*, 2010, vol. 4, n° 65, pp. 249- 256.

VERMERSCH Pierre, *L'entretien d'explicitation*, ESF Sciences Humaines., Psychologies et Psychothérapies, 2009, 208 p.

VINCK Dominique, « De l'objet intermédiaire à l'objet-frontière. Vers la prise en compte du travail d'équipement », *Revue d'anthropologie des connaissances*, 2009, vol. 3, n° 1, pp. 51- 72.

VINCK Dominique, « Les objets intermédiaires dans les réseaux de coopération scientifique. Contribution à la prise en compte des objets dans les dynamiques sociales », 1999, n° 40- 2, pp. 358- 414.

Annexe 11. Régression logistique et modèles utilisés

Principe de la régression logistique

La variable à expliquer `etat_fuyard` est une réponse binaire : 0 si le tronçon n'a pas été trouvé fuyard lors d'une inspection (i.e. réparé, du point de vue des données disponibles), 1 sinon. La variable n'est donc pas continue et ne peut faire l'objet d'une régression linéaire simple. Nous appliquons donc la théorie des « modèles linéaires généralisés » ("Generalized Linear Model", ou GLM), telle qu'exposée par l'ouvrage de référence de McCullagh and Nelder (1989).

On modélise alors la variable aléatoire y , qui est le nombre de tronçons fuyards parmi n réplicats indépendants qui partagent la même probabilité p d'être fuyard. Ces réplicats sont donc les tronçons de mêmes caractéristiques. y suit une distribution binomiale de paramètres (n, p) :

$$y \sim B(n, p) \Leftrightarrow P(y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}$$

La moyenne μ de y est liée à la combinaison linéaire des prédicteurs x_i par une fonction de lien g (transformation) :

$$g(\mu) = \eta = \beta_0 + \sum_{i=1}^m \beta_i x_i$$

Ici, nous utiliserons la transformation dite *logit* :

$$\eta = \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$\Leftrightarrow p = \frac{1}{1 + e^{-\eta}}$$

L'interprétation des coefficients (Estimate) est simple : la variable associée à un coefficient *coef* affecte la probabilité d'être fuyard d'un facteur e^{coef} (Odd-ratio). Donc si le coefficient est positif, l'effet de la variable est positif sur l'état fuyard, et inversement pour les coefficients négatifs. L'analyse de déviance permet, pour les modèles fixes, de calculer un effet global pour les facteurs qualitatifs. L'appel dans R se fait pour les modèles fixes via la commande `glm` : `glm(y ~ Var1 + Var2..., family = binomial(link = "logit"))`.

Pour les modèles mixtes (avec une/des variable(s) aléatoire(s)), l'appel se fait via la commande `glmer` du package `lme4` : `glmer(y ~ Var1 + Var2... + (1|Var aléatoire), family = binomial(link = "logit"))`.

Enfin, le test de la significativité d'un facteur aléatoire lié à l'intercept se fait avec la commande `glmmboot` du package `glmmML` : `glmmboot(y ~ Var1 + Var2..., family = binomial, cluster = Var aléatoire, boot = 1000)`.

Intérêt du modèle mixte

Le jeu de données utilisé est particulier à plusieurs titres, mais une particularité majeure est que les lignes (tronçons) ne sont pas indépendantes les unes des autres. En effet, on peut s'attendre à ce que l'état fuyard à une inspection d'un tronçon soit potentiellement plus corrélé :

- entre les différentes inspections du même tronçon qu'entre la même inspection de deux tronçons différents ;
- entre les tronçons d'une même cellule qu'entre les tronçons de deux cellules différentes ;
- entre les tronçons inspectés lors d'une même campagne qu'entre deux campagnes différentes.

La variation résiduelle de la réponse (non expliquée par les covariables) ne devrait donc pas être indépendante d'une observation à l'autre.

Pour pouvoir prendre en compte cette non-indépendance partielle des données, les GLM ont été étendus en GLM dits mixtes (GLMM), c'est à dire comprenant à la fois des effets fixes (les covariables que l'on veut étudier) et des effets dits aléatoires qui permettent de représenter la variation entre groupes. Brièvement, l'effet aléatoire permet aux coefficients du modèle de varier d'un groupe à l'autre, ici par exemple par cellule, cette variation suivant toujours une distribution normale.

L'idéal ici serait donc de créer un modèle de ce type : $glmer(etat_fuyard \sim Var1 + Var2... + (1|IDT/cell) + (1|IDcamp/cell), family = binomial(link = "logit"))$, avec les facteurs aléatoires étant IDT (ID tronçons) et IDcamp (ID de la campagne RF), niché dans le facteur cell (cellule) (modèle hiérarchique).

Cependant, calculer autant d'intercepts que de tronçons et de campagnes RF par cellule et faire converger ce modèle est impossible à cause de la taille trop importante du jeu de données. On se limite donc à un modèle plus simple qui prendra en compte uniquement l'effet aléatoire cellule.