Mise en place d'un modèle de fuite multi-états en secteur hydraulique partiellement instrumenté

Karim Claudio

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

PAR KARIM CLAUDIO

POUR OBTENIR LE GRADE DE **DOCTEUR**

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES ET CALCUL SCIENTIFIQUES

MAITRISE DES PERTES SUR LES RÉSEAUX D'EAU POTABLE

Mise en place d'un modèle de fuites multi-états en secteur hydraulique instrumenté

Soutenue le : 19 décembre 2014

Devant la commission d'examen formée de :

Rapporteurs	
Mme. Anne GÉGOUT-PETIT, Professeur	Université de Lorraine
Mme. Anne RUIZ-GAZEN, Professeur	Toulouse School of Economics
Examinateurs	
M. Vincent COUALLIER, Maître de Conférences (co-directeur de thèse)	Université de Bordeaux
Mme. Camelia GOGA, Maître de Conférences	Université de Bourgogne
M. Yves LE GAT, Ingénieur en chef	IRSTEA Bordeaux
M. Nikolaos LIMNIOS, Professeur	Université de Compiègne
M. Xavier LITRICO, Directeur du LyRE	Lyonnaise des Eaux
M. Jérôme SARACCO, Professeur (directeur de thèse)	Bordeaux INP-ENSC
Membres invités	
M. David DUCCINI, Directeur du Centre Technique Distribution	SUEZ ENVIRONNEMENT
M. Jean-François RENARD, Expert Réseau Distribution	Lyonnaise Des Eaux

Titre : Maîtrise des pertes sur les réseaux d'eau potable

Résumé L'évolution de l'équipement des réseaux d'eau potable a considérablement amélioré le pilotage de ces derniers. Le telérelevé des compteurs d'eau est sans doute la technologie qui a créé la plus grande avancée ces dernières années dans la gestion de l'eau, tant pour l'opérateur que pour l'usager. Cette technologie a permis de passer d'une information le plus souvent annuelle sur les consommations (suite à la relève manuelle des compteurs d'eau) à une information infra-journalière. Mais le télérelevé, aussi performant soit-il, a un inconvénient : son coût. L'instrumentation complète d'un réseau engendre des investissements que certains opérateurs ne peuvent se permettre. Ainsi la création d'un échantillon de compteurs à équiper permet d'estimer la consommation totale d'un réseau tout en minimisant les coûts d'investissement. Cet échantillon doit être construit de façon intelligente de sorte que l'imprécision liée à l'estimation ne nuise pas à l'évaluation des consommations. Une connaissance précise sur les consommations d'eau permet de quantifier les volumes perdus en réseau. Mais, même dans le cas d'une évaluation exacte des pertes, cela ne peut pas suffire à éliminer toutes les fuites sur le réseau. En effet, si le réseau de distribution d'eau potable est majoritairement enterré, donc invisible, il en va de même pour les fuites. Une fraction des fuites est invisible et même indétectable par les techniques actuelles de recherche de fuites, et donc irréparable. La construction d'un modèle de fuite multi-états permet de décomposer le débit de fuite suivant les différents stades d'apparition d'une fuite : invisible et indétectable, invisible mais détectable par la recherche de fuite et enfin visible en surface. Ce modèle, de type semi-markovien, prend en compte les contraintes opérationnelles, notamment le fait que nous disposons de données de panel. La décomposition du débit de fuite permet de fait une meilleure gestion du réseau en ciblant et adaptant les actions de lutte contre les fuites à mettre en place en fonction de l'état de dégradation du réseau.

Mots-clés : *Gestion des réseaux d'eau potable ; Échantillonnage par sondage stratifié ; Estimation des consommations ; Télérelevé des compteurs ; Décomposition du débit de fuite ; Modèle semi-markovien.*

Title : Mastering losses on drinking water network

Abstract The evolution of equipment on drinking water networks has considerably bettered the monitoring of these lasts. Automatic meter reading (AMR) is clearly the technology which has brought the major progress these last years in water management, as for the operator and the end-users. This technology has allowed passing from an annual information on water consumption (thanks to the manual meter reading) to an infra-daily information. But as efficient as AMR can be, it has one main inconvenient : its cost. A complete network instrumentation generates capital expenditures that some operators can't allowed themselves. The constitution of a sample of meters to equip enables then to estimate the network total consumption while minimizing the investments. This sample has to be built smartly so the inaccuracy of the estimator shouldn't be harmful to the consumption estimation. A precise knowledge on water consumption allows quantifying the water lost volumes on the network. But even an exact assessment of losses is still not enough to eliminate all the leaks on the network. Indeed, if the water distribution network is buried, and so invisible, so do the leaks. A fraction of leaks are invisible and even undetectable by the current technologies of leakage control, and so these leaks are un-reparable. The construction of a multi-state model enables us to decompose the leakage flow according to the different stages of appearance of a leak : invisible and undetectable, invisible but detectable with leakage control and finally detectable. This semi-Markovian model takes into account operational constrains, in particular the fact that we dispose of panel data. The leakage flow decomposition allows a better network monitoring but targeting and adapting the action of leakage reduction to set up according to the degradation state of the network.

Keywords : *Drinking water network management ; Stratified sample ; Water consumption estimation ; Automatic meter reading ; Leakage flow decomposition ; Semi-Markovian model*

Remerciements

Il y a tellement de personnes qui m'ont soutenu et aidé au cours de ces trois années ; vous m'avez tous à votre manière aidé à m'améliorer professionnellement mais aussi personnellement. Je ne sais pas s'il est possible de citer toutes ces personnes, d'avance je demande pardon à tous ceux que j'aurais pu oublier ici.

Je tiens tout d'abord à remercier encore une fois Mesdames Anne Gégout-Petit et Anne Ruiz-Gazen qui m'ont fait l'honneur de rapporter ce manuscrit de thèse. Je remercie aussi Mme Camelia Goga ainsi que M. Nikolas Liminios pour avoir accepté de faire partie de mon jury de thèse.

J'ai eu la chance de réaliser cette thèse chez Lyonnaise des Eaux, filiale de Suez Environnement. C'est ainsi que j'ai eu la chance de rencontrer Jean-François Renard, et la Direction Technique de Lyonnaise des Eaux, qui m'a beaucoup appris grâce à son expertise sur les réseaux d'eau potable. Je remercie également David Duccini, et son équipe, pour l'intérêt qu'il porte à mes travaux. Je tiens à vous remercier Jean-François et David d'avoir suivi et encadré cette thèse durant tout son déroulement et d'être une nouvelle fois présents pour cette étape finale qu'est la soutenance.

J'aimerais avoir une pensée pour l'équipe du Pôle Eau de l'entreprise régionale Bordeaux Guyenne, en particulier Michel Pic et Max Dubanchet pour leur connaissance du terrain, pour leur soutien et surtout pour avoir su donner un intérêt opérationnel à ces travaux de recherche.

Je remercie Xavier Litrico pour m'avoir donné la chance de réaliser ce travail de thèse au sein de son équipe au LyRE, pour son implication et sa volonté permanente de valoriser mon travail au sein du groupe et même au-delà.

Ces trois années de thèse CIFRE m'ont permis de rencontrer ma seconde "famille", je veux parler de l'équipe du LyRE. Bon ils sont 18 en tout alors pardonnez-moi mais je ne vais pas tous les citer même si cela pourrait paraître injuste d'oublier l'équipe de choc d'Ondeo Systems (je veux bien sûr parler de Jean-Jacques, Michaël, Jérôme, Martin et ses contrepèteries) ; Guillaume, Aude, Thibaud et AC qui forment la formidable équipe du CTP ou encore les écolos du LyRE, Damien et Thierry, menés d'une main de maître par Mélodie ! Non vraiment je n'ai pas le temps de remercier tout le monde, même s'il est impossible de parler du LyRE sans mentionner le pilier central de cette équipe : Véronique.

Et puis, ce serait outrancier de ne pas remercier mon acolyte au sein de la CEPA, Marie, qui doit me supporter tous les jours (et dont je dois en contrepartie supporter l'humour), mais aussi Julia, aka la Pink Chief !

Même si je ne peux remercier tout le monde, il m'est impossible d'oublier une certaine personne, le DuponT de mon DuponD, le Ooky de mon Wooky, le cartographe n°1 de la planète : Mr King Julien (merci d'être toujours à mes côtés, je me rends compte bien souvent qu'il n'est pas facile de travailler avec moi !).

Je tenais à terminer ces remerciements par les personnes qui m'ont accompagné tout au long de ces trois années : mes encadrants de thèse. Merci à Jérôme Saracco, mon directeur de thèse pour son implication et sa bienveillance tout au long de ces trois années. Merci aussi à Yves Le Gat, LE spécialiste des stats et de la dégradation des ouvrages. Il faut l'avouer, c'est réconfortant de travailler avec des personnes ayant une telle expertise statistique et opérationnelle. Un grand et chaleureux remerciement revient aussi à Vincent Couallier qui a été respectivement mon professeur, maître de stage et directeur de thèse, et qui a su mener ces trois rôles avec brio. Que tous les trois soient assurés de ma profonde gratitude.

Il y a parfois des rencontres qui bouleversent le cours d'une vie. C'est sans prétention que je peux affirmer qu'il m'est arrivé d'en vivre une. Je tiens ici à remercier une personne pour qui j'ai le plus profond respect (même si bien des fois les apparences laissent à penser le contraire), mon responsable au sein du LyRE, Cyril Leclerc. Je suis en train de vivre une aventure magnifique, qui n'en est qu'à son commencement, et c'est en grande partie à toi que je le dois.

Enfin, je remercie ma famille et mes amis qui sont toujours à mes côtés.

« Il y a trois sortes de mensonges : les mensonges, les gros mensonges et les statistiques ».

Benjamin Disraeli
(1804-1881)

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

L'eau a toujours été un enjeu majeur à travers le monde, mais les problématiques liées à cette ressource ont cependant évolué avec le temps. Dans un premier temps, les efforts se sont portés sur la distribution de l'eau à tous les habitants. Avec l'expansion des villes mais aussi des zones rurales, les réseaux de distribution d'eau potable se sont fortement agrandis. Aujourd'hui en France le réseau de distribution s'étend sur 906 000 km, desservant ainsi 24 millions d'abonnés [1]. On estime que 99% de la population est desservie en eau. Cet enjeu, majeur il y a quelques décennies, n'est plus aujourd'hui l'axe structurant la gestion des services d'eau potable en France.

La préoccupation des collectivités et des gestionnaires d'eau fut ensuite centrée sur la qualité de l'eau : la consommation domestique d'eau ne doit pas faire encourir un risque sanitaire. La directive 98/83/CE fixe au niveau européen des exigences à respecter au sujet de la qualité des eaux destinées à la consommation humaine. Cette directive a été transposée en droit français dans le code de la santé publique, aux articles R. 1321-1 à R. 1321-66. L'arrêté du 11 janvier 2007 fixe des normes de qualité à respecter pour un certain nombre de substances dans l'eau potable dont le chlore, le calcaire, le plomb, les nitrates, les pesticides et les bactéries.

Enfin, le problème actuel est la protection de la ressource, l'eau douce est une ressource limitée. "*La protection de la ressource du point de vue qualitatif comme quantitatif constitue donc un enjeu majeur des ministères en vue d'assurer une production pérenne d'eau potable*" [1]. L'eau prélevée à la ressource peut être divisée en deux composantes : les volumes consommés (à usage domestique et industriel) et les pertes dues en partie à la dégradation des réseaux de distribution. Le grand public a largement pris conscience de cette problématique, puisqu'une baisse des consommations est notable : la consommation d'eau potable des ménages s'établit en moyenne à 151 litres par jour et par habitant en 2008, contre 165 litres en 2004, soit une diminution de plus de 2% par an [1]. Cette baisse a été accélérée par l'émergence d'équipements mais aussi de comportement permettant d'économiser l'eau. Concernant les pertes, le volume d'eau prélevée n'arrivant pas jusqu'au robinet est estimé à 1.3 milliard de m^3, soit 22% de la production en 2008 [1]. La lutte contre les pertes d'eau potable en vue de protéger les ressources naturelles a poussé la ré-

1. source : Ministère de l'écologie, du développement durable et de l'énergie - Service de l'observation et des statistiques (SOeS), Enquête Eau 2008.

glementation à s'adapter ; en France, par exemple, la loi Grenelle II impose des contraintes sur la limitation des pertes sur les réseaux de distribution d'eau potable (différence entre le volume d'eau prélevée et l'eau consommé) et réglemente l'efficacité de ces réseaux (art. D. 213-48.14-1 et D. 213-74-1). De plus, au niveau local, les collectivités, déléguant la gestion du service d'eau potable aux entreprises privées, durcissent les contrats qu'elles engagent avec ces entreprises en imposant des contraintes tant au niveau des prélèvements d'eau à la ressource que des quantités d'eau potable perdues dans le réseau de distribution.

Ces différentes contraintes (environnementale, réglementaire et contractuelle) ont poussé les opérateurs à améliorer la gestion des réseaux d'eau potable. Les gestionnaires d'eau potable ont pour cela évalué l'efficacité de leurs réseaux à l'aide d'indicateur de performance. Pour clarifier la terminologie et uniformiser le calcul de ces indicateurs, l'Association Internationale de l'Eau (IWA) a établi une Balance de l'Eau [Alegre et al., 2000] présentée en Table 1.1. Cette balance décompose tout d'abord le volume délivré au réseau en deux sous-parties : les volumes qui ont été consommés et les volumes perdus. La balance permet de clairement définir cette notion de pertes et de distinguer les pertes commerciales des vols d'eau et des pertes liées au vieillissement du réseau.

TABLE 1.1 – Balance de l'eau définie par l'Association Internationale de l'Eau

<table>
<tr><td rowspan="8">Volume
délivré
au réseau</td><td rowspan="2">Consommation
autorisée</td><td>Conso. autorisée
facturée</td><td>Conso. facturée mesurée</td><td>Eaux
Facturées</td></tr>
<tr><td></td><td>Conso. facturée non mesurée</td><td></td></tr>
<tr><td rowspan="2">Conso. autorisée
non facturée</td><td>Conso. non facturée mesurée</td><td rowspan="6">Eaux
Non
Facturées</td></tr>
<tr><td>Conso. non facturée non mesurée</td></tr>
<tr><td rowspan="4">Pertes
en eau</td><td rowspan="2">Pertes
apparentes</td><td>Conso. non autorisée</td></tr>
<tr><td>Sous-comptage des compteurs</td></tr>
<tr><td rowspan="2">Pertes
réelles</td><td>Pertes sur les conduites
de distribution (canalisation)</td></tr>
<tr><td>Pertes sur les branchements
(jusqu'au poste de comptage)</td></tr>
</table>

Pour lutter contre les pertes sur les réseaux d'eau potable, l'IWA a identifé quatre pratiques permettant de réduire les volumes de pertes (voir [Lambert, 2002] ou [Farley and Trow, 2003]).

1. *La gestion patrimoniale ciblée* : cette pratique réunit l'ensemble des actions permettant de gérer de manière intelligente les installations (canalisations, branchements, compteurs, etc.) par le biais d'un plan de renouvellement adapté aux contraintes technico-économiques locales. Ces actions ont un effet sur le moyen/long terme (5 à 10 ans).

2. *La gestion de la pression* : il s'agit principalement de la modulation de pression. La technique consiste à isoler un secteur du réseau de distribution et à l'alimenter par un nombre limité de points d'entrée équipés de régulateurs de pression à partir desquels la pression est modulée. La pression ayant un impact sur les volumes d'eau

perdus ([Lambert, 2001]), les effets escomptés sont en premier lieu une réduction du taux de casse et donc du nombre de fuites, et en deuxième lieu une réduction du débit des fuites et donc du volume des pertes.

3. *La recherche active de fuites* : l'écoute du réseau enterré d'eau potable (grâce à diverses technologies comme les microphones d'écoute au sol) permet de localiser des fuites n'ayant pas encore "surfacé". La recherche active de fuites permet ainsi d'anticiper l'apparition en surface des fuites d'eau potable (voir par exemple [Pilcher, 2003]).

4. *La rapidité et la qualité d'intervention* : la durée d'écoulement d'une fuite est définie par trois temps : le temps de détection de la fuite, le temps de la localisation et le temps d'intervention. La réduction de ces trois temps permet de limiter ainsi la durée d'écoulement des fuites et de fait les volumes d'eau perdus. Ces trois temps dépendent essentiellement de l'opérateur et des moyens (humains et matériels) qu'il consacre à la bonne gestion de son réseau.

Afin d'opérationaliser ces quatre pratiques, les opérateurs ont instrumenté les réseaux d'eau potable, en commençant tout d'abord par les sectoriser. La sectorisation consiste à isoler des parties du réseau (appelées secteurs hydrauliques) en posant par exemple des vannes permettant de fermer le réseau de distribution. Au sein de ces secteurs, les quantités d'eau entrantes et sortantes (eaux transférées vers un autre secteur hydraulique) sont maîtrisées grâce notamment à la pose d'appareils de mesure comme les débitmètres [Morrison, 2004].
La mesure des volumes livrés au réseau par secteur hydraulique permet d'anticiper la détection des fuites grâce à la méthode du débit minimum nocturne (Minimum Night Flow, [Amoatey et al., 2014]). Le principe est le suivant : on calcule tout d'abord un débit entrant moyen (en m^3/h) sur une plage horaire située entre 00h et 04h T.U. (cette plage horaire est fixée par l'opérateur selon sa connaissance du réseau). Sur cette période de la journée, le débit est dit minimal puisqu'on suppose que les consommations sont elles-mêmes minimales et constantes sur cette période. Ainsi, toute augmentation du débit minimum nocturne est due à l'augmentation des pertes et donc à l'apparition et la dégradation d'une fuite. La décomposition du débit minimum nocturne peut être approfondie (voir par exemple [Fantozzi and Lambert, 2010]), mais nous ne nous attarderons que sur le premier stade de décomposition. Cette méthode, même si elle s'est avérée efficace, comporte certaines limites. Il n'est pas possible de calculer directement le volume de pertes journalières. De plus, la méthode fait l'hypothèse d'une constance des consommations nocturnes au cours du temps, ce qui peut s'avérer inexact en présence de consommateurs industriels sur le secteur.

L'instrumentation des réseaux se traduit ensuite par la pose de régulateurs de pression sur un secteur hydraulique. Grâce à ces régulateurs, la pression de l'ensemble du secteur est modulée heure par heure à la valeur minimale permettant d'assurer la qualité de service satisfaisante. Aux heures de pointe, les régulateurs sont totalement ouverts et la pression livrée n'est pas réduite.

Enfin, parallèlement et indépendamment de la gestion des pertes, les compteurs d'eau potable des usagers ont été équipés d'émetteurs télérelevés. Cette technologie récolte et retransmet automatiquement les index de consommation (consommation cumulée) sur un pas de temps variant entre une et six heures. L'information, auparavant acquise grâce aux relèves manuelles sur un pas de temps variant entre un mois et un an, est maintenant disponible automatiquement grâce au télérelevé sur un pas de temps infra-journalier (une heure ou six heures selon le paramétrage des émetteurs télérelevés). Cependant, le télérelevé a été principalement développé et valorisé par rapport aux bénéfices qu'il apportait en termes de facturation (relève plus précise, facturation au réel, détection précoce des fuites chez l'usager, etc.) mais à l'heure actuelle, son exploitation à des fins opérationnelles n'en est qu'à son commencement.

Une plus-value de l'utilisation des données issues du télérelevé des compteurs d'eau potable est de connaître les volumes consommés sur un secteur hydraulique à un pas de temps journalier voire infra-journalier. Une meilleure connaissance des consommations induit une meilleure connaissance des pertes sur les réseaux d'eau potable (voir Table 1.1). Si les différents indicateurs de performance sont calculés de manière plus précise (meilleure qualité des données d'entrée, pas de temps plus fins), il est alors possible de réduire les temps d'écoulement des fuites en anticipant la détection des fuites.

Cependant les calculs des pertes sur les réseaux d'eau potable, comme pour les autres indicateurs de performance, nécessitent de connaître la consommation de tout un secteur hydraulique et donc d'équiper en télérelevé l'intégralité des compteurs d'eau de ce secteur. L'équipement exhaustif des compteurs d'eau peut s'avérer trop long dans certains cas voire trop coûteux dans d'autres. Il est ainsi important de pouvoir calculer précisément la consommation globale en s'affranchissant de ces contraintes, soit en planifiant l'équipement progressif des compteurs, soit en identifiant un échantillon stratégique de compteurs à équiper.

Enfin, le calcul des pertes à lui seul ne permet pas d'optimiser la gestion du réseau. En effet, le concept BABE, pour Burst And Background Estimates ([Lambert, 1994]), permet de différencier les fuites selon trois états (voir la Figure 1.1) : les *fuites diffuses* (fuites non reportées et non détectables par recherche de fuites), les *fuites détectables* (fuites non reportées mais détectables par recherche de fuites) et les *casses manifestes* (fuites reportées n'ayant pas encore été réparées). En fonction du type de fuites, les moyens d'intervention diffèrent (renouvellement, recherche de fuite ou réparation). Ainsi l'optimisation de la gestion du réseau suppose préalablement la décomposition du débit de fuite.

L'objectif principal de cette thèse est la construction d'un outil (méthodologie) permettant de détecter "en temps réel" (à J+1) l'apparition de fuites détectables sur le réseau et d'estimer la part de fuites diffuses dans le débit de fuite. Pour cela, l'outil se basera sur les données issues du télérelevé. Il convient de pouvoir exploiter ces données tout en limitant les investissements nécessaires à l'équipement des compteurs.

Nous nous attarderons ainsi dans une première partie sur une méthode d'échantillonnage permettant d'identifier les compteurs à équiper en télérelevé tout en minimisant les coûts d'instrumentation. Une information fiable sur la consommation d'un secteur a de grandes implications sur la gestion du réseau. En particulier, l'estimateur de la consommation de la population doit être suffisamment précis afin de permettre la détection de

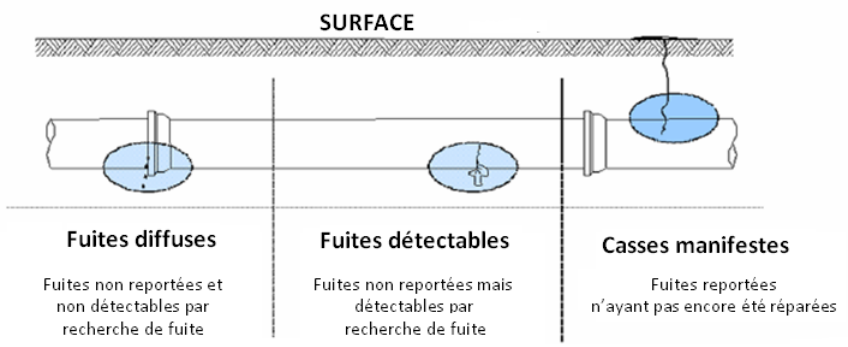


FIGURE 1.1 – le concept BABE : décomposition des fuites en trois états (schéma issu de [Fantozzi and Lambert, 2010])

fuites potentielles. En conséquence, l'échantillon à constituer (taille d'échantillon, sélection des individus) doit conduire à un estimateur du total dont la précision est contrôlée. L'application de la théorie des sondages en population finie répond à cette problématique.

S'il n'existe pas dans la littérature en théorie des sondages de méthodes complètes pour constituer un échantillon dans un cas appliqué comme celui-ci, les différentes étapes permettant la construction d'un panel d'usagers à télérelever y sont développées : la construction des strates (*e.g.* [Dalenius and Hodges, 1959] ou [Serfling, 1968]), le taux optimal d'échantillonnage ([Tillé, 2001]) ou encore l'allocation de l'échantillon dans chaque strate ([Cochran, 1977], [Ardilly, 2006]). A partir de ces différents travaux, il est alors possible de définir un échantillon permettant d'estimer, de manière fiable, la consommation en eau d'une population. Une fois l'échantillon constitué, des informations auxiliaires peuvent être disponibles. Ainsi, si la précision de l'estimateur initial ne permet pas de détecter de manière suffisamment précise les fuites sur le réseau d'eau potable, ces informations vont nous permettre de redresser l'estimateur, afin d'en améliorer potentiellement la précision.

Cet estimateur des consommations totales nous permettra alors de construire la première version de la balance de l'eau (Table 1.1) en calculant, à partir des volumes délivrés au réseau, les volumes des pertes en eau.

Nous consacrerons ensuite une seconde partie à la mise en place d'un modèle de décomposition des fuites selon le concept BABE, à partir de l'estimation des pertes tirées des données de consommations télérelevées. Ce modèle est basé sur l'hypothèse qu'une fuite passe naturellement par trois états de dégradation. Des premiers travaux ([Chesneau, 2006]) ont permis de construire des modèles de décomposition des volumes perdus mais ces travaux sont limités par les données utilisées en entrée (la consommation est calculée à partir de données agrégées et estimées sur un pas de temps hebdomadaire) mais aussi par les hypothèses faites permettant la convergence des modèles. Le modèle que nous construirons tire son origine des processus semi-markoviens et doit permettre de modéliser non pas les fuites mais la dégradation des ouvrages d'eau potable (canalisations et branchements) selon les fuites qu'ils subissent. Une des limites à notre application est que les individus observés ne sont pas suivis en continu ; nous disposons cependant d'inspections ponctuelles non régulièrement réparties dans le temps. L'application d'un modèle

semi-markovien sur des données de panel nous permettra de modéliser l'évolution de la dégradation d'un ouvrage selon les trois états : l'ouvrage ne subit aucune fuite ou uniquement des fuites diffuses, l'ouvrage subit des fuites détectables mais invisibles en surface et enfin l'ouvrage subit des casses manifestes. Il sera alors possible pour chaque ouvrage de calculer la probabilité qu'il soit dans un des trois états précédents. Enfin le croisement de cette probabilité et de l'estimation des pertes (obtenue grâce à l'estimation des consommations) permettra de décomposer les pertes globales selon les trois états. Cette décomposition du débit de fuite permettra d'une part d'approfondir la connaissance sur les pertes sur le réseau de distribution d'eau mais aussi d'optimiser la gestion du réseau en adaptant les moyens d'intervention.

Chapitre 2

Données et cadre d'étude

Les données exploitées dans le cadre de cette thèse serviront essentiellement au calage des modèles construits. Nous distinguons deux types de données : les données patrimoniales et les données métrologiques. Les données patrimoniales permettent de caractériser le réseau d'eau potable. Les données métrologiques quant à elles sont dynamiques ; elles proviennent de l'exploitation quotidienne du réseau et nous pouvons considérer qu'elles sont fournies en temps continu.

Les données sont issues de contrats gérés par Lyonnaise des Eaux, précisément ceux de l'Entreprise Régionale Bordeaux Guyenne. Nous nous intéressons à deux contrats en particulier, chacun ayant une plus-value vis-à-vis de la disponibilité et la qualité des données en présence : le contrat de concession d'eau potable de la Communauté Urbaine de Bordeaux (CUB) et le contrat d'affermage d'eau potable de Canéjan (*cf.* Annexe A).

2.1 Le patrimoine d'eau potable

Le patrimoine d'eau potable décrit l'ensemble des ouvrages permettant l'adduction de l'eau potable, c'est-à-dire le transport et la distribution de l'eau potable. L'eau est puisée dans une **ressource** naturelle (rivières, nappes souterraines, etc.) puis acheminée vers une **usine de traitement** où a lieu la phase de potabilisation ; l'eau en sort traitée et viable à la consommation. Elle passe ensuite par un ensemble de conduites de transport pour être stockée dans un **réservoir** ou un **château d'eau**. Finalement, l'eau transite via le réseau de distribution : des **canalisations** vers les **branchements** puis des branchements vers les **compteurs** d'eau, pour arriver chez l'usager (voir Figure 2.1).

Nous porterons notre attention sur le réseau de distribution c'est-à-dire depuis le point de stockage jusqu'au compteur. Les pertes réelles sur le réseau de distribution décrites à la Table 1.1 ont lieu principalement sur les canalisations et les branchements, ce sont donc ces deux types d'ouvrages que nous traiterons par la suite. Ces ouvrages ne sont pas homogènes en composition ; il existe différents types de matériau, chacun ayant ses propres caractéristiques, ses propres forces, ses propres faiblesses.

Le modèle multi-états que nous tentons de construire nous permettra de mieux comprendre et d'anticiper la dégradation de ce patrimoine.

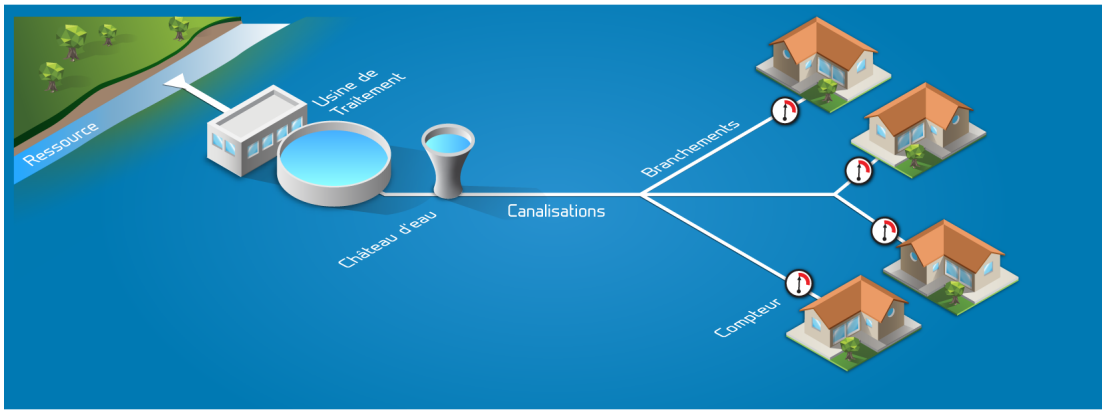


FIGURE 2.1 – Schéma du réseau d'adduction d'eau potable.

2.1.1 Les données relatives au patrimoine d'eau potable

Il existe une source d'information principale concernant le patrimoine d'eau potable : le Système d'Information Géographique (SIG). Le SIG de Lyonnaise des Eaux, nommé *MMS-APIC*, géo-référence, de manière plus ou moins exhaustive, les différents ouvrages d'eau potable et d'assainissement.

Les caractéristiques structurelles du patrimoine

a - Les canalisations

Les canalisations sont les principales conduites de distribution d'eau potable, elles acheminent l'eau du point de stockage jusqu'aux branchements. Il s'agit d'ouvrages dont la longueur graphique dans le SIG peut varier entre quelques mètres et quelques centaines de mètres. Les canalisations d'eau potable ne sont pas toutes faites du même matériau, celui-ci ayant évolué avec le temps. Au milieu du XIX^{ème} siècle, les réseaux étaient principalement faits d'acier ou de fonte appelée fonte grise alors qu'au début du XX^{ème}, on voit apparaître des conduites en béton ou en amiante ciment. Une nouvelle génération de fonte, la fonte ductile, est ensuite apparue dans les années 60. Enfin dans les années 50, les premières conduites en plastique (PVC ou polyéthylène) sont installées sur le réseau de distribution. La Figure 2.2 illustre la répartition des canalisations en fonction du matériau, pour les contrats de la CUB et de Canéjan. Étant donné la forte hétérogénéité des longueurs des canalisations relevées dans le SIG, le réseau de distribution est évalué en termes de linéaire et non pas en nombre de canalisations.

La majorité du réseau de distribution de la CUB est composée de fonte (certaines canalisations en fonte grise datant de la fin du XIX^{ème} siècle sont toujours en service) alors que sur le réseau de Canéjan, le plastique (en particulier le PVC) est prédominant.

Le décret d'application de la loi Grenelle II du 27 janvier 2012 impose de nouvelles obligations en matière de description des réseaux d'eau potable. Les réseaux doivent faire l'objet d'un descriptif détaillé concernant le linéaire des canalisations, les dates ou périodes de pose, les matériaux et les diamètres des canalisations. Ainsi ce décret incite les opérateurs à améliorer le niveau de connaissance du patrimoine concernant les canalisa-

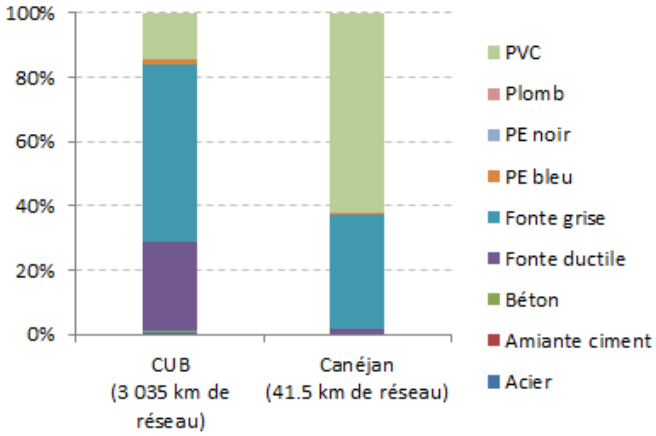
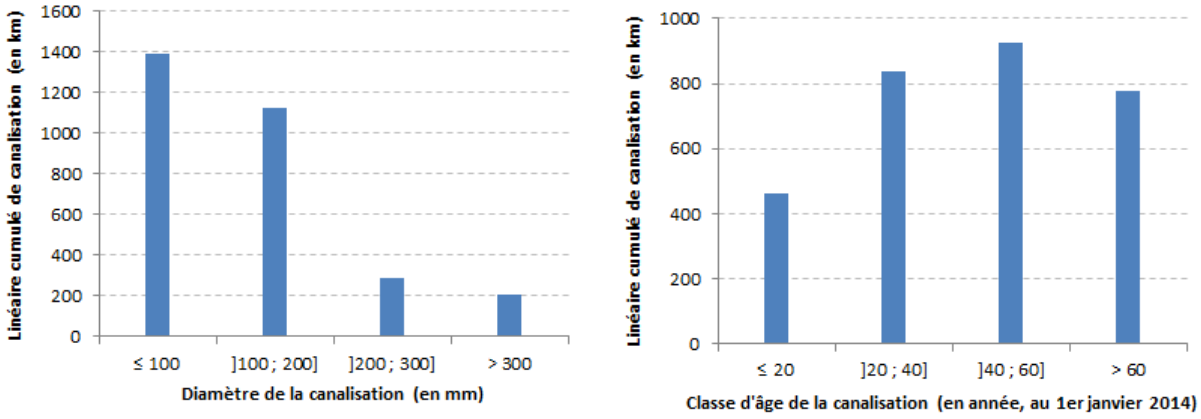


FIGURE 2.2 – Répartition (en %) du linéaire de canalisations en fonction du matériau pour les réseaux de la CUB et de Canéjan (source SIG *MMS-APIC*, 2014).

tions d'eau potable. A ce jour, le niveau de connaissance du patrimoine "canalisations" est plus élevé sur le réseau de la CUB que sur celui de Canéjan (seulement 20% des canalisations sont datées sur le réseau de Canéjan). Les principales données disponibles sur les canalisations du réseau de la CUB par l'interrogation du SIG sont les suivantes :

- le matériau,
- la date de pose,
- la longueur,
- le diamètre,
- la position géographique.

Les Figures 2.3 présentent la répartition du linéaire de canalisation selon (a) les classes de diamètre et (b) d'âge de la canalisation en 2014 sur le réseau de la CUB.



(a) (b)

FIGURE 2.3 – Répartition du linéaire de canalisation par (a) classe de diamètre et (b) classe d'âge au 1er janvier 2014 (source *MMS-APIC*, 2014)

b - Les branchements Les branchements sont les tuyaux reliant les canalisations aux compteurs d'eau. Ils sont plus nombreux que les canalisations, mais d'un linéaire plus

faible et de plus petits diamètres. La législation, imposant une connaissance profonde sur le réseau d'eau potable et principalement sur les canalisations de distribution, n'implique pas directement les branchements d'eau potable. De fait les données sur les branchements (localisation, date de pose et matériau) ne sont souvent que partiellement disponibles et nous pouvons fortement douter des données concernant les longueurs individuelles et leur diamètre.

Cependant, selon un engagement contractuel, l'ensemble des branchements d'eau potable de la CUB doit être géo-référencé. Concernant l'information sur le matériau et la date de pose, l'ensemble des branchements posés après le 1er janvier 2006 sont renseignés exhaustivement ; un travail de reconstitution des données est à réaliser pour les branchements posés avant cette date. Cependant, il n'existe pas de tel engagement sur le contrat de Canéjan, la disponibilité des données concernant les branchements est donc fortement limitée, comme le montre le Tableau 2.1.

TABLE 2.1 – Nombre de branchements installés et géo-référencés.

Contrat	Nb branchements installés (source RAD* 2012)	Nb branchements géo-référencés (source SIG *MMS-APIC*)	% de branchements géo-référencés
CUB	204 350	201 853	98.8%
Canéjan	2 023	877	43.3%

*Rapport Annuel du Délégataire

On constate ainsi que, sur le réseau de la CUB, 98.8% des branchements sont géo-référencés dans le SIG alors que seuls 43.3% le sont sur Canéjan. Les branchements non géo-référencés ne sont pas exploitables (aucune information sur leur matériau, leur date de pose ou encore sur les fuites qu'ils ont pu subir). Ainsi dans l'exploitation des données patrimoniales, nous nous focaliserons sur les données de la CUB. Il faut noter que les ouvrages non géo-référencés pour la CUB, considérés comme des branchements dans le SIG, sont en réalité des accessoires (poteaux incendie, bouches à clé, etc.) qui ne seront pas modélisés.

Compte tenu de la différence de connaissance patrimoniale que nous avons entre les réseaux de la CUB et de Canéjan, les modèles que nous construirons seront basés sur les données patrimoniales de la CUB. La Figure 2.4 montre la répartition des branchements encore en service sur la CUB selon le type de matériau.

L'environnement et le fonctionnement du patrimoine

Si la structure d'un ouvrage définit en grande partie ses conditions de vieillissement, l'environnement et le fonctionnement de cet ouvrage auront un impact significatif sur l'évolution de son état au cours du temps. Différents facteurs environnementaux et fonctionnels peuvent interagir avec les canalisations, nous ne les listerons pas exhaustivement mais nous citerons uniquement ceux pour lesquels l'exploitation et le croisement avec les bases de données patrimoniales ont été possibles :

– *La nature des sols* : les sols argileux peuvent fragiliser les ouvrages. [Morris, 1967] et [Clark, 1971] soulignent le fait qu'il s'agit en réalité du retrait-gonflement des

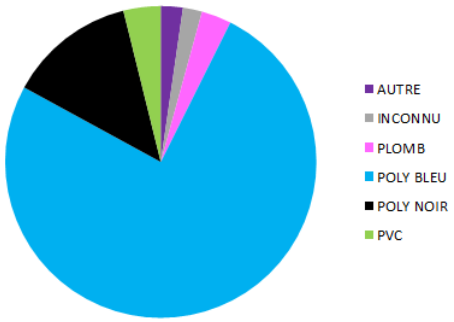


FIGURE 2.4 – Répartition par matériau du nombre de branchements (source *MMS-APIC*, 2014)

argiles qui contribue à la dégradation des ouvrages. Nous utiliserons alors une variable indicatrice de la pose d'un ouvrage dans un sol argileux. Par ailleurs les sols corrosifs peuvent eux aussi altérer le bon fonctionnement de la canalisation notamment pour celles en métal, fonte grise, fonte ductile ou acier. La corrosion d'un sol peut être approchée à partir de sa résistivité.

– *Le climat* : l'effet de ce facteur est complexe et souvent indirect. On suppose très fortement que le froid a un effet sur les conduites en métal. [Kleiner and Rajani, 2002] ont observé que des pics de casses survenaient périodiquement à la fin de l'hiver quand la température des sols était la plus basse. Il faut aussi noter que de longues périodes de temps sec ont tendance à favoriser le retrait-gonflement des argiles [Wols and Van Thienen, 2014], les conduites en plastique étant plus sensibles à ce genre de phénomène.

– *Le trafic* : la présence d'un fort trafic routier au dessus d'un ouvrage peut créer des vibrations du sol provoquant une dégradation accélérée de la canalisation.

– *La densité de branchements d'une canalisation* : plus le nombre de branchements connectés à la canalisation est élevé, plus celle-ci présente des points de fragilité.

– *La pression moyenne du réseau* : La pression a un effet direct sur l'évolution des fuites (détail en sous-section 2.2.2). La pression moyenne du réseau est une pression calculée à partir d'un modèle hydraulique qui permet d'estimer la pression en tout point du réseau.

Les interventions sur le patrimoine

Afin d'étudier le vieillissement des ouvrages d'eau potable (branchements et canalisations), il est nécessaire de récolter en plus des données sur leur structure, leur fonctionnement et leur environnement, des données relatives aux interventions sur ces ouvrages. Les interventions que nous considérons ici concernent en premier lieu les réparations survenues soit après constatation (de l'opérateur ou d'un particulier) car les fuites étaient visibles en surface, soit après une enquête de l'opérateur sur le réseau pour localiser une fuite n'ayant pas encore "surfacé". Les réparations sur le réseau sont enregistrées dans le

SIG ce qui permet d'associer à une intervention l'ouvrage impacté. Il faut noter que toutes les fuites ne sont pas conservées pour les modèles que nous allons construire : en effet certaines casses ont été provoquées par des entreprises tierces (lors de travaux), ce qui ne traduit pas en réalité un vieillissement du réseau mais plutôt un accident. Les casses provoquées par les entreprises tierces seront supprimées de notre analyse.

Une autre information pertinente pour nos modèles concerne le renouvellement du patrimoine d'eau potable (canalisation et branchement). Le renouvellement d'un ouvrage par un nouvel ouvrage, plus "jeune" et donc dans un meilleur état, a un impact sur les pertes en eau, notamment par la réduction des fuites non visibles en surface.

2.1.2 Reconstitution des données patrimoniales

Au vu des données disponibles, nous pouvons considérer que les données du patrimoine "canalisation" de la CUB sont exhaustives concernant la localisation, la date de pose, le matériau, la longueur graphique ou encore le diamètre. Concernant les branchements, si ceux-ci sont géo-référencés de façon quasi-exhaustive sur la CUB, il n'est pas garanti que toutes les données soient disponibles, en particulier les dates de pose. Il est nécessaire, pour les besoins de la modélisation, de reconstituer les dates de pose afin d'exploiter le maximum de données possible. Afin de reconstituer les dates de pose manquantes, nous nous servirons du croisement de la base de données branchements avec différentes bases de données métier afin d'attribuer à un branchement une date de pose en cohérence avec son matériau. Les différentes sources de données exploitées sont les suivantes :

- la base d'intervention *G2* qui recense notamment la création de branchements neufs,
- la base de données Clientèle qui recense la pose des premiers compteurs,
- la base des lotissements qui recense la construction des lotissements et notamment des branchements posés afin de raccorder les habitations au système de distribution d'eau potable,
- la base de données patrimoniales des canalisations issues de *MMS-APIC*.

Ces différentes bases de données ont permis de rattacher un branchement à une intervention (création d'un branchement) ou un élément du réseau (compteur, lotissement, canalisation) qui nous a permis d'estimer les dates de pose. Afin de limiter les erreurs d'estimation, il convient de s'assurer de la cohérence entre la date de pose estimée et le matériau du branchement. Les règles suivantes sont donc utilisées pour valider une date de pose estimée :

- si le branchement est en polyéthylène (PE) noir, alors l'année de pose doit être comprise entre 1970 et 1985,
- si le branchement est en polyéthylène (PE) bleu, alors l'année de pose doit être supérieure ou égale à 1986,
- si le branchement est d'un autre matériau, alors la date est validée (les autres matériaux ne sont pas modélisés).

Récapitulatif sur la complémentation des dates de pose

Le tableau 2.2 résume les résultats obtenus par les différentes sources sur la complémentation des dates de pose des branchements.

TABLE 2.2 – Complémentation des dates de pose des branchements

Source Date de Pose	Nombre
MMS-APIC*	73 457
G2	2 133
Canalisation	25 833
1er compteur	22 341
Lotissement	9 427

**renseignée initialement*

Ainsi au final, nous avons 72% de la base de données branchements datée. Sur les 28% restants, une majorité concerne les branchements en PE Noir, comme le montre la Figure 2.5. Cette fraction de branchements non datés, même si elle représente une part non négligeable de la base branchements, ne sera pas modélisée à cause du manque d'information "précise" que l'on a sur ces branchements.

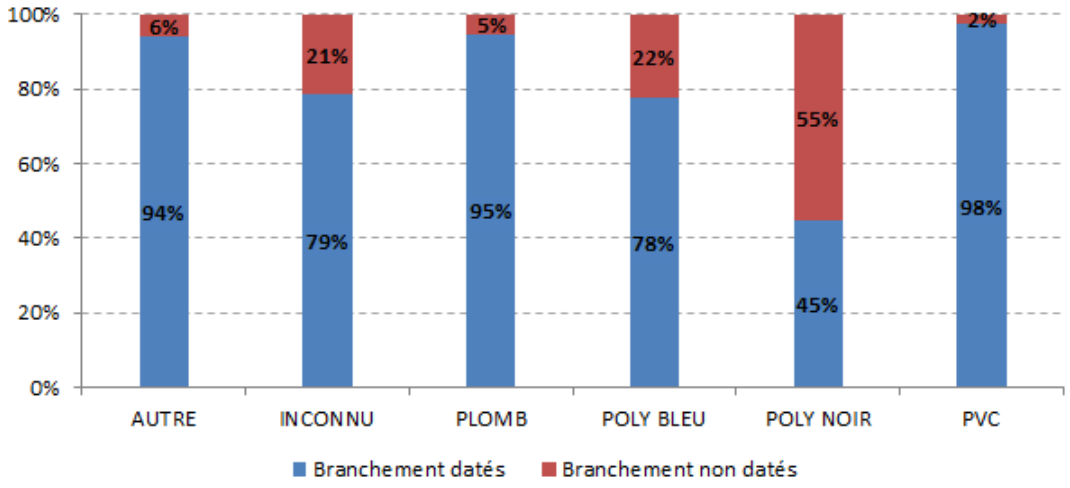


FIGURE 2.5 – Datation finale des branchements par nature de matériau

2.2 L'instrumentation du réseau

Les données métrologiques recensent les données issues des différents capteurs posés sur le réseau. Nous exploiterons trois types de données métrologiques dans notre étude :

- les volumes d'eau délivrés au réseau,
- les pressions sur le réseau,
- les volumes d'eau consommés.

Les données métrologiques sont le reflet de l'exploitation quotidienne du réseau. Le pas de temps de ces trois données est assez fin (entre 5 min et 6h) pour considérer qu'elles sont disponibles en temps continu et en temps réel (en réalité les données sont disponibles à $J+1$). Les différentes données ainsi collectées permettent un suivi "en temps réel" du réseau et facilitent la détection d'anomalies.

2.2.1 Les données de volumes livrés au réseau

Les débitmètres de sectorisation mesurent sur un pas de temps très fin (5 min) les volumes livrés au réseau (VLAR), comme on peut le voir à la Figure 2.6 et au Tableau 2.3. Si on considère un secteur hydraulique simple, avec un seul point d'entrée d'eau et uniquement pour sortie les consommations, il est facile de calculer les pertes sur un réseau, connaissant les volumes entrants et consommés.

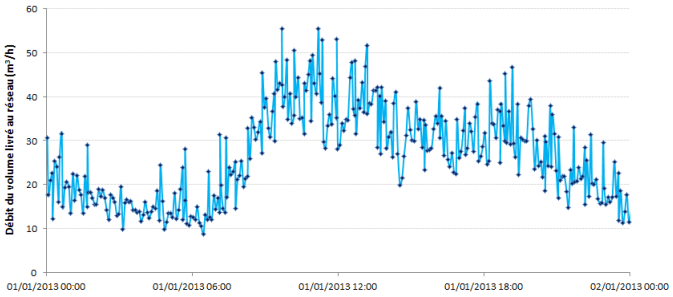


FIGURE 2.6 – Débits de volumes livrés

TABLE 2.3 – Données des débits de volumes livrés au réseau

Date	Débit (m3/h)
01/01/2013 00 :04	17.7
01/01/2013 00 :09	20.96
01/01/2013 00 :14	22.53
01/01/2013 00 :19	25.29
01/01/2013 00 :24	24.14
01/01/2013 00 :29	16.08
01/01/2013 00 :34	26.31

Les données fournies par les bases de données internes sont des débits (volumes rapportés à une durée). Si entre deux instants t_1 et t_2, un débit de q m^3/h a été mesuré, alors le volume v correspondant vaut :

$$v = q \times (t_2 - t_1)$$

où la différence $t_2 - t_1$ est exprimée en heure si le débit q est en m^3/h.

Les données de débit entrant permettent non seulement de connaître les volumes mis en service mais également de calculer le débit minimum nocturne. Cette technique permet de détecter l'apparition des fuites sur le réseau. La méthode consiste à calculer un débit moyen la nuit sur une plage horaire située entre 00h et 04h T.U. ; la définition exacte variant suivant le secteur [Alkasseh et al., 2013]. L'hypothèse est faite que, durant cette plage horaire, les consommations sont minimales et constantes (on suppose que les consommations domestiques sont nulles et que seules restent les consommations industrielles, constantes dans le temps). Ainsi toute hausse du débit minimum nocturne ne peut être imputée qu'à l'apparition de fuites sur le réseau. Nous verrons plus tard les limites de cette méthode, notamment sur l'hypothèse des consommations nocturnes constantes.

2.2.2 Les données de pression

Parallèlement aux données de pressions moyennes du réseau disponibles grâce au calage d'un modèle hydraulique, l'instrumentation du réseau à l'aide de capteurs de pression

permet de connaître la pression de l'eau dans le réseau, sur un point fixe du réseau, de manière continue (voir Figure 2.7). La pression a un effet capital sur l'évolution des fuites. [Van Zyl and Clayton, 2007] décrivent une relation entre la pression et le débit de fuite, que d'autres auteurs comme [Lambert, 2001] traduiront plus simplement : la variation du débit de fuite évolue comme la variation de la pression à une puissance N_1.

$$L_1/L_0 = (P_1/P_0)^{N_1} \tag{2.1}$$

où L_i et P_i sont respectivement le débit de fuite et la pression à un instant i.

FIGURE 2.7 – Exemple de chronique de données de pression d'eau

2.2.3 Le télérelevé des compteurs d'eau

Le fonctionnement du service

Le télérelevé des compteurs d'eau est un dispositif permettant d'obtenir en temps réel et à distance différentes informations, la principale étant la consommation d'eau d'un usager. Le service remplace la relève manuelle du compteur faite le plus souvent une seule fois par an. Le dispositif est illustré à la Figure 2.8 : Une tête émettrice, posée sur le compteur (1), envoie un signal vers une antenne réceptrice placée à proximité (2), cette dernière renvoyant à un serveur (3) tous les index qu'elle a pu acquérir sur les 24 dernières heures.

Pour les contrats gérés par Lyonnaise des Eaux, la technologie développée par Ondeo Systems (filiale du Groupe Suez Environnement) permet de récolter des index de consommations toutes les heures ou toutes les six heures (paramétrage par défaut des émetteurs).

Données disponibles

Le télérelevé permet d'acquérir les index de consommation d'un usager. Il s'agit en réalité de sa consommation cumulée depuis l'installation du compteur. La différence d'index entre deux instants donnés permet de connaître la consommation sur cette période de temps. Le dispositif de télérelève renvoie les index de consommation sur un pas de temps qui peut être soit d'une heure (24 informations par jour) soit de six heures (4 informations par jour). Un exemple de données récoltées sur un pas de temps horaire est montré au

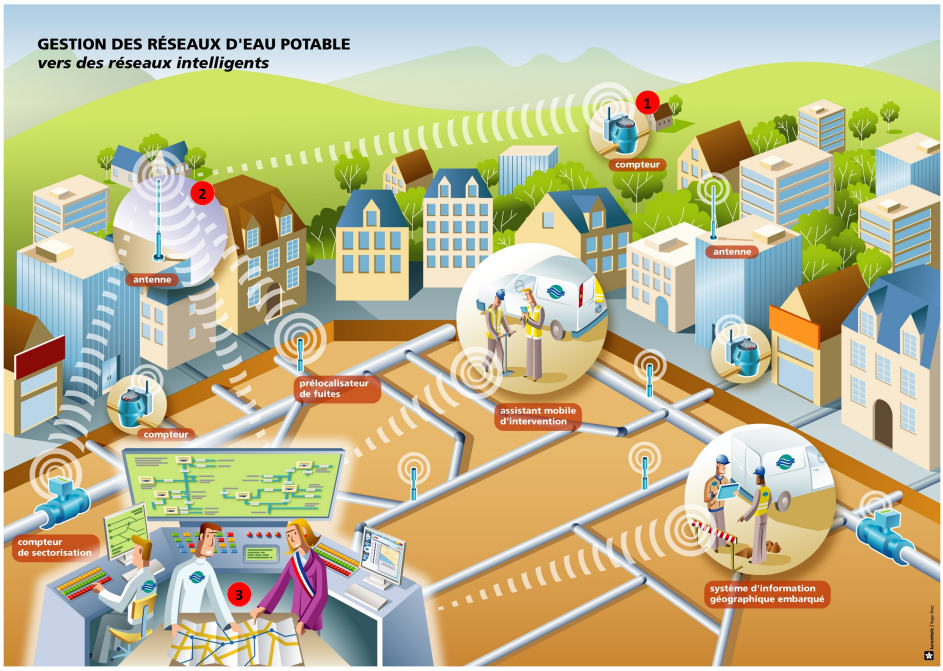


FIGURE 2.8 – Le fonctionnement du télérelevé des compteurs d'eau

Tableau 2.4.

TABLE 2.4 – Exemple de télé-transmission des index par le télérelevé.

Date de télé-transmission	Index de consommation (en m^3)
04/02/2014 20 :40 :01	1209.125
04/02/2014 19 :40 :01	1208.974
04/02/2014 18 :39 :58	1208.916
04/02/2014 17 :39 :58	1208.916
04/02/2014 16 :39 :58	1208.915
04/02/2014 15 :39 :58	1208.915
04/02/2014 14 :39 :51	1208.915

Avantages et inconvénients du service

La relève automatique des compteurs d'eau présente certes de nombreux avantages pour l'usager tout comme pour l'exploitant. Cependant à l'instar des "smart grids", les compteurs électriques ou gaz intelligents, elle comporte un inconvénient d'un point de vue financier ou encore des limites d'un point de vue éthique (respect de la vie privée de l'usager) .

La communication autour des compteurs intelligents met en avant les avantages pour l'utilisateur. Pour une majorité d'usagers (les ménages en particulier), la facturation se basait jusque là sur une relève annuelle des index, avec une estimation des consommations à mi-année, corrigée par la nouvelle relève 6 mois plus tard. La récolte automatique

des index entraîne alors une facturation plus précise, basée sur une consommation réelle. Cette information infra-annuelle apporte à l'usager une meilleure connaissance sur ses consommations, permettant ainsi à l'usager de mieux maîtriser sa consommation en eau potable. Enfin, la connaissance horaire ou journalière des consommations alerte l'usager en cas de consommations anormales en particulier une surconsommation (basée sur un référentiel qu'il aura défini au préalable) ou en cas de fuite dans le domaine privé (entre le compteur d'eau et le robinet).

Mais si l'usager a une connaissance plus précise de ses consommations, il en va de même pour l'exploitant. La Commission Nationale de l'Informatique et des Libertés (CNIL) met en garde les utilisateurs contre certaines dérives liées à l'utilisation des compteurs communicants (eau, gaz et électricité) dans un article paru le 24 janvier 2013 [1] :

> *« Le principal risque provient d'une nouvelle fonctionnalité offerte par les compteurs communicants : la courbe de [consommation]. Cette courbe de [consommation] est constituée d'un relevé, à intervalles réguliers (le "pas de mesure"), de la consommation [...] de l'abonné. Plus le "pas de mesure" est faible, plus les mesures sur une journée sont nombreuses et fines, permettant d'avoir des informations précises sur les habitudes de vie des personnes concernées [permettant] notamment d'identifier les heures de lever et de coucher, les périodes d'absence, etc. ».*

Enfin un aspect non négligeable du service est son financement. Les coûts de pose des émetteurs et des antennes réceptrices sont directement répercutés sur le prix de l'eau (après décision de la collectivité), et les éventuels services associés sont à la charge (et au choix) de l'usager. Ainsi le financement du télérelevé est entièrement à la charge de l'usager, directement ou indirectement.

Seulement le télérelevé n'est pas seulement avantageux pour l'usager mais aussi pour la collectivité et l'exploitant. Tout d'abord, si l'usager n'a plus besoin d'être présent à son domicile pour un relevé des compteurs, il n'est plus nécessaire pour l'exploitant de lancer des campagnes de relevés sur le territoire. La facturation des consommations, basée sur des données mesurées et non plus estimées, conduit à une diminution des réclamations clientèles dues à une mauvaise estimation des consommations. Enfin, le fait de proposer un suivi journalier (ou infra-journalier) régulier des consommations est un support de base à la sensibilisation des usagers sur leur consommation et l'adoption de gestes hydro-économes.
Mais si le télérelevé améliore la relation avec l'usager, il permet aussi d'optimiser le pilotage du réseau. La construction d'une balance de l'eau (voir Table 1.1) sur une échelle infra-annuelle permet un suivi plus précis des pertes en eau et une meilleure maîtrise du rendement de réseau. Le pilotage de l'exploitation en temps réel permet de suivre de près la performance opérationnelle et les impacts environnementaux des différentes politiques de gestion (interventions sur le réseau, sensibilisation des usagers).

1. http ://www.cnil.fr/linstitution/actualite/article/article/compteurs-communicants-premieres-recommandations-de-la-cnil/

Déploiement du télérelevé en France

Après l'électricité et le gaz, aujourd'hui les compteurs d'eau deviennent eux aussi des compteurs communicants. Si le service n'est pas autant déployé que pour les autres énergies, on estime qu'en 2016 plus de 30 millions de compteurs d'eau seront télérelevés à travers le monde (source : *Pike Research*). Les trois grands opérateurs d'eau français (Lyonnaise des Eaux, Véolia Eau et Saur) proposent l'équipement des compteurs domestiques et professionnels. En France, la solution d'Ondeo Systems est contractualisée sur près de 1.1 millions de compteurs. Lyonnaise des Eaux a développé ce service depuis 2006 sur plus de 200 contrats.

Au niveau de l'Entreprise Régionale Bordeaux Guyenne, les déploiements contractuels ont débuté en 2009 d'abord sur la commune de Canéjan puis se sont étendus à différents syndicats comme celui de Carbon Blanc devenu, depuis fin 2013, la première collectivité de France en nombre de compteurs d'eau connectés [2].

Parallèlement aux déploiements contractuels, certains compteurs de la communauté urbaine de Bordeaux ont été équipés dans le cadre de projet de recherche. Lancé en 2007, un projet d'estimation du rendement de réseau journalier a été mis en place par l'Entreprise Régionale Bordeaux Guyenne de Lyonnaise des Eaux. Pour cela, un secteur hydraulique, nommé "Cote 50", a été sélectionné : il s'agit d'une zone de 17 km de réseau, comportant environ 1600 compteurs d'eau potable. Dans le cadre de ce projet, 10% des compteurs ont été équipés en télérelevé horaire, ce taux ayant été fixé suivant des contraintes économiques. L'échantillon formé a été établi à partir d'un sondage stratifié selon la consommation annuelle individuelle : la population a été découpée en 8 strates de consommation, définies selon des références communément utilisées dans le domaine de l'eau (par exemple, la borne supérieure de la strate 1 correspond à la consommation annuelle d'un foyer d'une personne).

Le Tableau 2.5 récapitule les zones à partir desquelles nous pourrons exploiter les données télérelevées.

TABLE 2.5 – Liste des zones d'étude télérelevées exploitables

Zone	Nb de compteurs	Nb de compteurs télérelevés	Taux d'instrumentation	Fréquence de télé-transmission
CUB (Cote 50)	1604	147	9.5%	1h
Canéjan	2069	2015	97%	6h

Ainsi, pour la première partie des travaux menés, une attention toute particulière est portée sur le secteur de Canéjan. En effet, l'utilisation d'un secteur entièrement équipé permettra de tester la fiabilité d'une méthode d'échantillonnage et d'estimation des consommations, en confrontant les résultats obtenus aux données mesurées.

Traitement des données télérelevées

Le télérelevé des compteurs fournit une information riche sur les consommations d'eau. Cependant, il est parfois possible que des problèmes surviennent lors de la télé-

2. *http ://www.sudouestnumerique.net/index.php/smallnews/detail ?newsId=15257*

transmission des index (provoquant, par exemple, des plages manquantes d'index), soit lors de l'envoi de l'index vers le récepteur, soit lors du stockage sur le serveur. La Figure 2.9 montre un exemple d'erreur de télé-transmission qui nécessite un traitement des données.

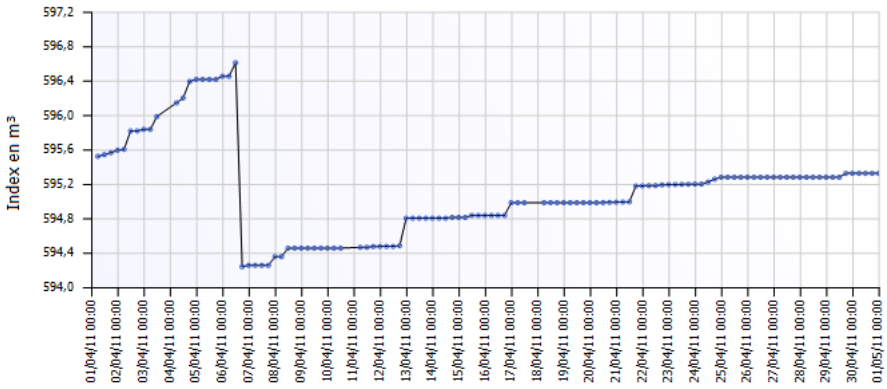


FIGURE 2.9 – Exemple d'erreur de télé-transmission d'un index de consommation

Un index de consommation représentant la consommation cumulée, il ne peut être que croissant ou constant, en aucun cas décroissant comme illustré en Figure 2.9. Par ailleurs pour éviter la saturation des antennes réceptrices, les émetteurs envoient les index de consommation en décalé. Ainsi, à partir des données brutes, nous ne disposons pas de l'information sur la consommation au même instant pour tous les individus, ce qui entraîne des difficultés pour réaliser un bilan de la consommation à un instant t donné. Pour pallier ce problème, un recalage des index à heure fixe doit être réalisé pour tous les compteurs d'un secteur considéré.

Il apparaît ainsi nécessaire qu'un traitement des données brutes doit être réalisé avant toute exploitation de ces données. Ce traitement se décompose en trois parties : tout d'abord l'identification (et la suppression) des index aberrants et des données manquantes, puis la reconstruction des valeurs rejetées ou inconnues et enfin un recalibrage des données à heure fixe.

Pour des raisons de confidentialité, cette partie n'est pas disponible.
Pour plus d'informations, merci de contacter l'auteur :
karim.claudio@lyonnaise-des-eaux.fr

2.3 Conclusion du chapitre

La robustesse des modèles statistiques vient en grande partie de la fiabilité des données d'entrée. Il convient alors, préalablement à toute étape de modélisation, de valider les données d'entrée en s'assurant de leur exhaustivité mais aussi de leur exactitude. Les modèles que nous tentons de construire ici sont complexes d'une part par les phénomènes que l'on tente de modéliser mais aussi du fait des données exploitées. Nous exploiterons

par la suite des données statiques, liées aux caractéristiques des ouvrages patrimoniaux, mais aussi des données dynamiques, liées à l'exploitation des réseaux. La multiplicité des données implique inévitablement la multiplicité des sources d'erreur. Il convient alors de choisir comme secteur d'étude celui qui propose la plus grande fiabilité des données mais aussi la meilleure disponibilité de ces données. Au regard de ce dernier critère, notre analyse est ainsi limitée au secteur de l'entreprise régionale Bordeaux Guyenne. Concernant la fiabilité des données, les bases de données du contrat de Bordeaux fournissent des données robustes et pouvant potentiellement être complétées au besoin concernant le patrimoine. L'équipement du réseau bordelais avec des débitmètres de sectorisation et des capteurs de pression permet de récolter en temps continu des données essentielles à l'exploitation des réseaux. Ce secteur sera donc particulièrement intéressant pour mettre en place le modèle de décomposition du débit de fuite. Cependant, une lacune majeure de ce secteur est le manque d'information concernant les consommations. En effet, le faible équipement des compteurs d'eau en émetteurs télérelevé ne permet pas d'assurer un calcul fiable des consommations totales d'un secteur hydraulique. Un échantillon expérimental a été constitué sur le secteur de la Cote 50 (voir carte en Figure 2.10) pour permettre à partir de 10% de la population d'estimer la consommation totale du secteur. Il n'y a aucun moyen de fiabiliser l'échantillon ou en tout cas d'évaluer sa fiabilité. C'est pourquoi une partie des travaux menés porte sur la construction d'un échantillon permettant d'estimer la consommation totale d'un secteur en vue d'évaluer les pertes sur ce même secteur. Cette méthodologie d'échantillonnage devra être calée et validée sur un secteur dont l'intégralité des compteurs est télérelevée afin de pouvoir confronter les résultats obtenus aux données mesurées. Le secteur de Canéjan semble donc tout à fait approprié comme secteur d'étude à la phase d'échantillonnage et d'estimation des consommations que nous développerons au chapitre 3.

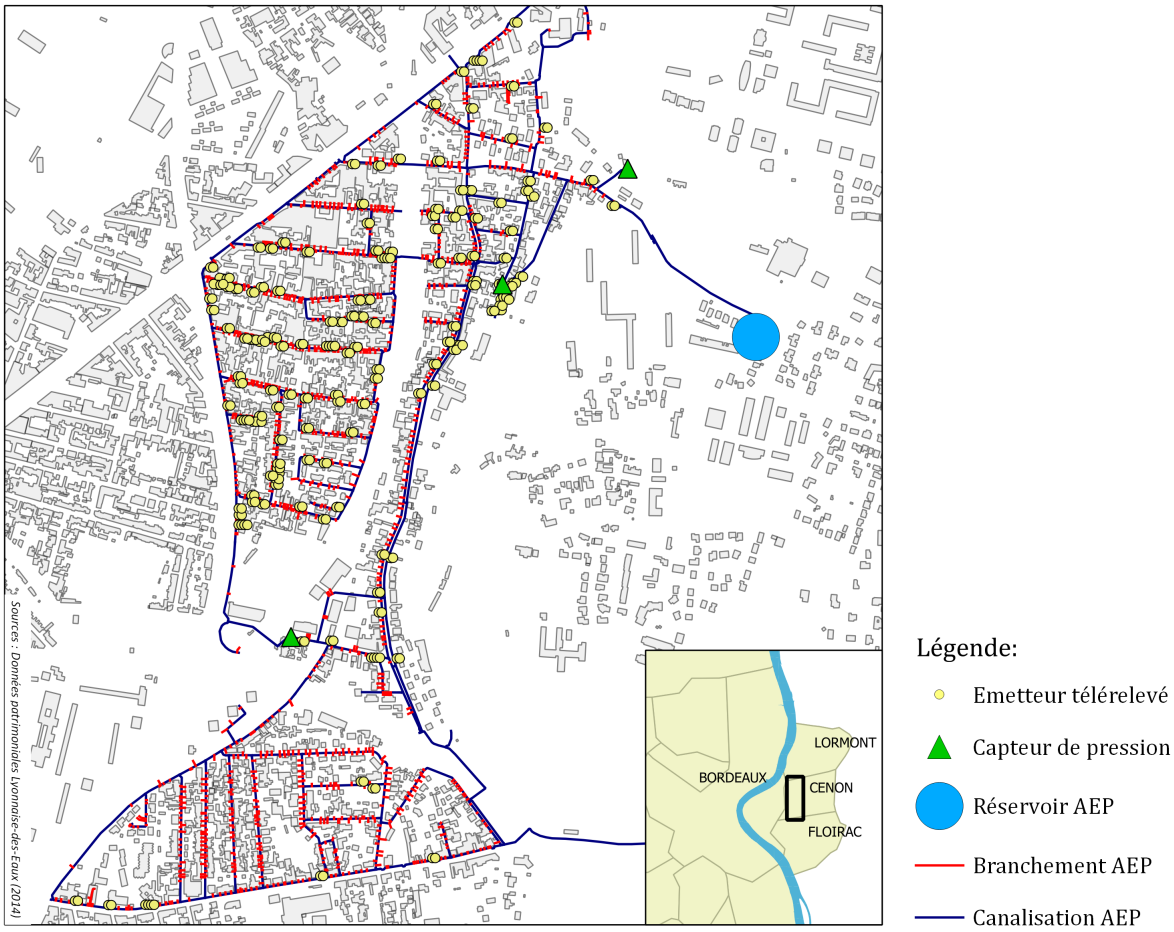


FIGURE 2.10 – Le réseau d'alimentation en eau potable (AEP) de la Cote 50 à Cenon

Chapitre 3

Estimation de la consommation d'eau à partir d'un échantillon d'usagers télérelevés

Préambule

Ce chapitre est consacré à l'élaboration d'un plan d'échantillonnage permettant d'estimer de manière fiable (nous définirons clairement par la suite la notion de précision de l'estimateur) la consommation totale d'une population. Il convient, pour évaluer l'efficacité de notre sondage, de pouvoir comparer les résultats d'estimation aux vraies données de consommations mesurées. Dans cette optique, nous avons choisi comme terrain d'étude la commune de Canéjan (33). Cette commune, comptant 1954 compteurs d'eau, est entièrement télérelevée depuis la fin de l'année 2009 (à l'heure actuelle 2015 compteurs sont équipés mais seuls les 1954 sélectionnés dans ce chapitre ont une profondeur d'historique de données dépassant deux ans). Nous élaborerons spécifiquement notre échantillon en vue d'estimer la consommation journalière ou hebdomadaire totale sur l'année 2011, puis nous verrons la performance de notre échantillon sur une période postérieure. L'individu statistique considéré ici est donc le compteur d'eau.

3.1 Constitution d'un échantillon et estimation du total

Notons notre population U, de taille N, dont on souhaite extraire un échantillon s, de taille n. Cet échantillon, une fois constitué et équipé en télérelevé, doit permettre de récolter des données sur la consommation en eau potable des usagers. La variable d'intérêt est la consommation individuelle (notée Y_i, $i \in [\![1,N]\!]$) sur un pas de temps défini (le jour ou la semaine). Nous nous intéressons à son total $T_Y = \sum_{i \in U} Y_i$.
L'aspect temporel de la variable d'intérêt conduit à considérer que l'estimation du total peut être faite à chaque pas de temps. La variable d'intérêt sera donc notée $Y_i(t)$, la consommation de l'individu i au temps t, dont nous voulons estimer le total $T_Y(t)$.

Au vu des données disponibles et dans la mesure où l'objectif est d'obtenir une information infra-annuelle, le choix se porte sur un pas de temps soit journalier soit hebdomadaire. Pour permettre une détection plus rapide des fuites et une diminution des durées d'écoulement, notre choix se portera alors sur un pas de temps journalier, plus approprié

pour cette problématique. Nous appelons $\mathscr{T}$ le nombre de jours sur la période d'étude considérée ; nous allons donc estimer $\mathscr{T}$ consommations journalières totales.

L'estimateur utilisé pour calculer la consommation journalière totale est l'estimateur par sondage stratifié. Le principe est de diviser la population en groupes d'individus homogènes (appelés "strates") au sein desquels on extrait un échantillon. Nous utiliserons par la suite les notations suivantes : la population est divisée en H strates de consommation, notées G_h, de taille N_h (avec $\sum_h N_h = N$). L'intersection de la strate G_h et de l'échantillon s forme un sous-échantillon g_h, d'effectif n_h (avec $\sum_h n_h = n$). L'estimateur du total des consommations journalières se calcule de la manière suivante :

$$\hat{T}_Y(t) = \sum_{i \in s} \frac{1}{\pi_i} Y_i(t) = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in g_h} Y_i(t) = \sum_{h=1}^{H} N_h \hat{\bar{y}}_h(t), \tag{3.1}$$

avec $\pi_i = \frac{n_h}{N_h}$ si $i \in G_h$.

Pour pouvoir utiliser cet estimateur, il convient au préalable de stratifier la consommation à l'aide d'une variable, appelée "variable de stratification" qui doit être exhaustive et corrélée à notre variable d'intérêt pour obtenir un estimateur performant. Historiquement, le rélevé manuel des compteurs permet de connaître chaque année la quantité d'eau consommée par chaque individu de la population durant l'année qui vient de s'écouler. Il semble raisonnable de réaliser un sondage stratifié à partir de cette variable connue exhaustivement, cette donnée étant très fortement corrélée à notre variable d'intérêt comme le montre la Figure 3.1.

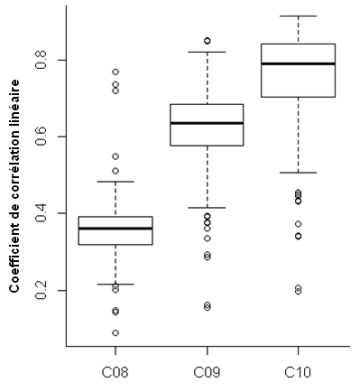
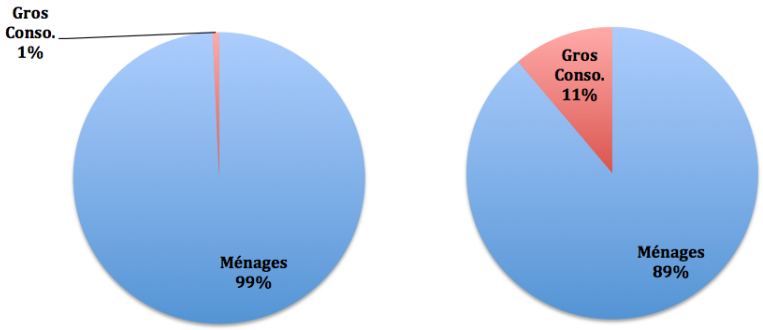


FIGURE 3.1 – Boîtes à moustaches des $\mathscr{T}$ coefficients de corrélation linéaire entre les consommations journalières individuelles en 2011 et les consommations annuelles individuelles de 2008, 2009 et 2010

On remarque que plus la variable de stratification est proche, temporellement parlant, de la variable d'intérêt, plus la corrélation entre les deux variables est forte. Pour une estimation de la consommation durant l'année A, la variable de stratification la plus adéquate (et disponible) est la consommation annuelle individuelle en $A-1$. Nous noterons par la suite X la variable de stratification : la consommation annuelle individuelle de l'année $A-1$.

3.1.1 Stratification de la population

Dans une population, on distingue deux types d'usagers : les ménages et les gros consommateurs. Cette dernière catégorie de la population (essentiellement composée d'industriels et d'institutions) peut se définir, selon les experts métier, comme les usagers dont la consommation annuelle individuelle est supérieure ou égale à 1000 m^3. Ce groupe de consommateurs, malgré son faible effectif (ils sont environ 1% de la population) représente une part importante des consommations annuelles totales (en moyenne 10%). Cette asymétrie entre les effectifs et les consommations s'observe à la Figure 3.2.



(a) (b)

FIGURE 3.2 – Effectif des ménages et gros consommateurs (a) et consommations annuelles (b) en 2010

De plus, les gros consommateurs ont un comportement totalement hétérogène entre eux (leur consommation allant de 1000 à 4000 m^3/an en 2010). Afin de limiter l'influence de ces valeurs extrêmes, nous attribuons à ces individus un poids de sondage égal à 1 ([Beaumont et al., 2013]) : ils sont donc regroupés dans une même strate enquêtée exhaustivement.

On se fixe pour des raisons opérationnelles un nombre H de strates à créer. La borne des gros consommateurs étant imposée, il reste $H^* = H - 1$ strates à définir. Le choix du nombre de strates se base sur l'homogénéité des strates. Pour quantifier cette homogénéité, nous avons choisi comme indicateur la somme pondérée des variances intra-strates de la variable de stratification X ($V_W(X)$) :

$$V_W(X) = \sum_{h=1}^{H^*} \frac{N_h}{N} S_{Xh}^2 \tag{3.2}$$

où

- N_h est l'effectif de la strate h,
- S_{Xh}^2 est la variance de X au sein de la strate h.

La Figure 3.3 montre l'évolution de $V_W(X)$ en fonction du nombre de strates.

Tel que cela est spécifié dans [Kpedekpo, 1973] le gain en efficacité, bien que viable pour une augmentation initiale du nombre de strates, devient marginal après un certain point. Il peut y avoir des situations où une augmentation du nombre de strates peut même

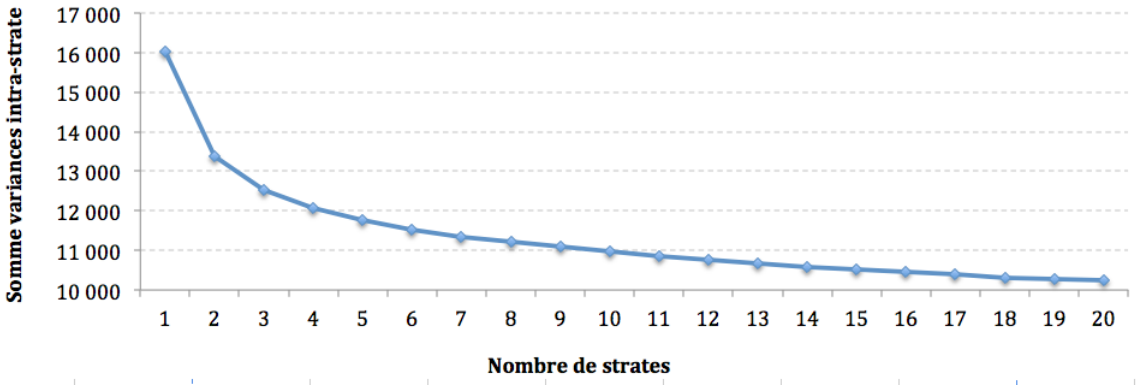


FIGURE 3.3 – $V_W(X)$ en fonction du nombre de strates

conduire à diminuer l'homogénéité des strates et de fait réduire l'efficacité de l'échantillonnage stratifié. Nous avons donc choisi le nombre de strates de telle sorte que l'ajout d'une nouvelle strate n'apporte qu'un gain marginal dans l'homogénéité des strates. Ceci est mesuré par :

$$H^* = \arg\max_h \left\{ \frac{V_W(X|h-1) - V_W(X|h)}{V_W(X|h-1)} \geq 1\% \right\}, \tag{3.3}$$

où $V_W(X|h)$ est la somme pondérée des variances intra-strate en considérant une stratification en h strates (hors gros consommateurs). Le seuil choisi (1%) est purement arbitraire mais permet d'obtenir un nombre de strates suffisamment important pour discriminer efficacement la population tout en limitant le nombre de strates.

La Table 3.1 montre l'évolution de la somme pondérée des variances intra-strates et du ratio associé en fonction du nombre de strates.

TABLE 3.1 – Choix du nombre de strates en fonction de l'évolution de $V_W(X)$

h	1	2	$\cdots$	8	9	10	11
$V_W(X\|h)$	16 019	13 372	$\cdots$	11 209	11 088	10 972	10 869
$\frac{V_W(X\|h-1) - V_W(X\|h)}{V_W(X\|h-1)}$		16.50%	$\cdots$	1.20%	1.10%	1.10%	0.90%

Au vu de cette table, nous allons donc former 10 strates en plus de celle des gros consommateurs. Le découpage optimal de la population est celui qui permettrait d'obtenir des strates homogènes, c'est-à-dire obtenir une variance minimale dans chaque strate. D'après Dalenius ([Dalenius, 1950]), cela revient à trouver x_1,..., x_{H^*} solutions du système suivant

$$\frac{S_{Xh}^2 + (x_h - \bar{X}_h)^2}{S_{Xh}} = \frac{S_{Xh+1}^2 + (x_{h+1} - \bar{X}_{h+1})^2}{S_{Xh+1}}, \qquad \forall h \in [\![1, H^* - 1]\!]$$

où $\bar{X}_h$ est la moyenne de la variable X au sein de la strate h. S'il n'existe pas de solutions analytiques à ce problème, plusieurs approximations ont été proposées (voir par exemple [Serfling, 1968], [Nicolini, 2001] et [Lavallée and Hidiroglou, 1988]), nous retenons toutefois la méthode de Dalenius et Hodges ([Dalenius and Hodges, 1959]), pour

sa simplicité d'implémentation.

Soit f_X la densité de X. La quantité $\tilde{x}_h$ est une approximation de x_h, si $\tilde{x}_h$ est racine de l'équation suivante :

$$\int_{-\infty}^{\tilde{x}_h} \sqrt{f_X(t)}dt = \frac{h}{H^*} \times \int_{-\infty}^{+\infty} \sqrt{f_X(t)}dt.$$

Dans la majorité des applications, la vraie densité f_X est inconnue. Cochran [1977] propose alors une approche pratique basée sur la distribution empirique de la variable X. Notons C_X la fonction définie $\forall \xi \in \mathbb{N}^*$ par :

$$C_X(\xi) = \sum_{x=1}^{\xi} \sqrt{\sum_{i=1}^{N} \mathbb{1}_{[x-1<X_i\leq x]}}. \tag{3.4}$$

Cochran préconise de choisir la borne supérieure x_h de la strate h de telle sorte que

$$x_h = \arg\max_{\xi} \{C_X(\xi) \leq \frac{h}{H^*} C_X(X_M)\} \tag{3.5}$$

où X_M est la valeur maximale de X (dans ce cas précis, X_M = 1000, car au delà les individus sont dans une strate déjà définie). Dans notre application, $C_X(X_M) = 655$, d'après l'équation (3.5), les seuils à atteindre sont alors 65.5 ($= (655/10) \times 1$), 131, 196.5, 262, 327.5, 393, 458.5, 524 et 589.5. La Table 3.2 montre un exemple de sélection de la borne supérieure des premières strates.

TABLE 3.2 – Exemple de sélection des bornes supérieures des strates

ξ (en m^3)	$C_X(\xi)$	Strate
1	5.47	1
2	8.30	1
...	...	...
29	62.51	1
30	**64.96**	1
31	67.79	2
...	...	...
49	**128.93**	2
50	132.67	3
...	...	...
999	655	10

Ainsi la borne supérieure de la strate 1 est 30 m^3 et celle de la strate 2 est 49 m^3. Les bornes finales des strates de consommation sont présentées en Table 3.3.

TABLE 3.3 – Stratification de la population

strate	Bornes	N_h
1	[0 ; 30 [	192
2	[30 ; 49 [	155
3	[49 ; 65 [	184
4	[65 ; 79 [	213
5	[79 ; 94 [	206
6	[94 ; 109 [	207
7	[109 ; 129 [	203
8	[129 ; 150 [	205
9	[150 ; 185 [	182
10	[185 ; 1000 [	193
11	[1000 ; +∞ [	14

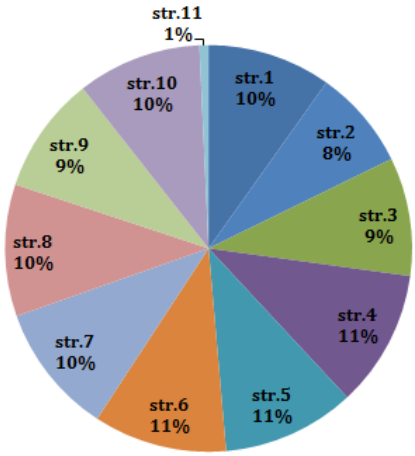


FIGURE 3.4 – Répartition de la population par strate

On constate au vu de la répartition de la population dans chaque strate (Figure 3.4) que, mise à part la strate des "gros consommateurs", la population se répartit de façon homogène dans chacune des dix premières strates. La phase de stratification est à présent achevée, nous avons défini nos strates : leur nombre H ainsi que leurs bornes. Notons que pour réaliser ce découpage, les deux étapes (choix du nombre et des bornes) ne peuvent se faire de façon dissociée ; il faut réaliser le processus de choix du nombre en calculant à chaque fois les bornes au moyen de l'équation (3.5). Il faut maintenant sélectionner notre échantillon et le répartir à travers nos différentes strates.

3.1.2 Taille de l'échantillon et répartition dans les strates

La taille de l'échantillon a deux impacts conséquents sur le sondage : d'une part le nombre de compteurs à équiper définit les investissements nécessaires à la mise en place de l'échantillon et d'autre part la taille de l'échantillon joue sur la précision de l'estimateur. La Figure 3.5 montre l'évolution de l'écart-type de l'estimateur stratifié en fonction du taux de sondage.

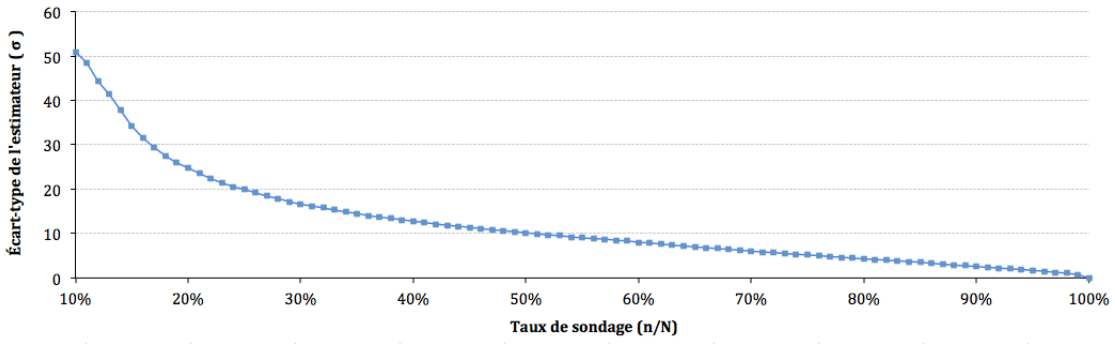


FIGURE 3.5 – Ecart-type de l'estimateur en fonction du taux de sondage

La variance de l'estimateur de Horvitz-Thomspon du total (dans le cadre d'un sondage

stratifié) s'écrit :

$$V(\hat{T}_Y(t)) = \sum_{h=1}^{H} N_h^2 \frac{(1-f_h)}{n_h} S_{Yh}^2(t), \tag{3.6}$$

où

- $S_{Yh}^2(t)$ est la dispersion de la variable $Y(t)$ au sein de la strate h,
- n_h est l'effectif de l'échantillon dans la strate h,
- $f_h = n_h/N_h$ est le taux de sondage de la strate h.

Afin de simplifier la lecture des équations, nous omettrons par la suite l'indice de temps t. L'objectif est de minimiser la taille d'échantillon n tout en atteignant un objectif de variance de l'estimateur égale à σ^2. Nous pouvons noter que la strate H, la strate des "gros consommateurs" a été définie de telle sorte que tous les individus au sein de celle-ci soient enquêtés. De fait nous cherchons à déterminer la taille d'échantillon n^* telle que :

$$n = n^* + N_H. \tag{3.7}$$

Au vu de l'équation (3.6), si la strate H est entièrement enquêtée, alors elle n'aura aucun poids dans le calcul de la variance de l'estimateur. En partant de l'écriture de la variance, il est possible d'établir une relation entre la taille de l'échantillon n^* et l'écart-type σ :

$$\begin{aligned}
\sigma^2 &= \sum_{h=1}^{H^*} N_h^2 \left(\frac{1-f_h}{n_h}\right) S_{Yh}^2 \\
&= \sum_{h=1}^{H^*} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{Yh}^2 \\
&= \sum_{h=1}^{H^*} \frac{N_h^2}{n_h} S_{Yh}^2 - \sum_{h=1}^{H^*} N_h S_{Yh}^2
\end{aligned} \tag{3.8}$$

L'équation (3.8) fait intervenir une quantité inconnue n_h (nombre d'individus sondés dans la strate h) que nous transformons en

$$n_h = n^* \times a_h \tag{3.9}$$

où a_h est l'allocation de l'échantillon dans la strate h. En injectant (3.9) dans (3.8), nous obtenons :

$$\begin{aligned}
\sigma^2 = \sum_{h=1}^{H^*} \frac{N_h^2}{a_h n^*} S_{Yh}^2 - \sum_{h=1}^{H^*} N_h S_{Yh}^2 &\Leftrightarrow \frac{1}{n^*} \sum_{h=1}^{H^*} \frac{N_h^2}{a_h} S_{Yh}^2 = \sigma^2 + \sum_{h=1}^{H^*} N_h S_{Yh}^2 \\
&\Leftrightarrow \frac{1}{n^*} = \frac{\sigma^2 + \sum_{h=1}^{H^*} N_h S_{Yh}^2}{\sum_{h=1}^{H^*} \frac{N_h^2}{a_h} S_{Yh}^2} \\
&\Leftrightarrow n^* = \frac{\sum_{h=1}^{H^*} \frac{N_h^2}{a_h} S_{Yh}^2}{\sigma^2 + \sum_{h=1}^{H^*} N_h S_{Yh}^2}
\end{aligned} \tag{3.10}$$

Ainsi d'après l'égalité (3.10), il est possible, en fonction d'une précision σ souhaitée, de déterminer le nombre n adéquat d'individus à sonder. Cependant, cette égalité fait intervenir S_{Yh}^2 où

$$S_{Yh}^2 = \frac{1}{N_h - 1} \sum_{i \in G_h} (Y_i - \overline{Y}_h)^2$$

quantité inconnue car elle fait intervenir la variable d'intérêt $Y_i(t)$. Il n'est pas non plus possible de l'estimer par

$$s_{Yh}^2 = \frac{1}{n_h - 1} \sum_{i \in g_h} (Y_i - \overline{Y}_h)^2$$

car pour cela il faudrait que l'échantillon soit déjà constitué et nous cherchons justement à déterminer la taille de l'échantillon. Nous pourrions contourner ce problème en utilisant les variances S_{Xh}^2.Cependant, cette option pose ici différents problèmes :

- il est difficile de transcrire l'erreur tolérée à un pas de temps journalier vers un pas de temps annuel,
- l'exploitation de données annuelles ne permet pas de mettre en évidence la saisonnalité des données et notamment la forte hétérogénéité des consommations en périodes estivales,
- l'application de la formule (3.10) avec des données annuelles n'a pas renvoyé des résultats satisfaisants.

Comme il est recommandé dans Fellegi [2010], le calcul de n est "difficile à obtenir et une approximation est fréquemment faite à partir de populations similaires". Un calcul préliminaire sur des secteurs entièrement télérelevés pourra permettre de calculer la taille d'échantillon nécessaire pour obtenir une précision σ requise par des experts métier et de créer un panel de contrats de référence utilisable pour d'autres applications. Ainsi, dans notre application, nous utiliserons les données de consommation journalière sur Canéjan pour déterminer une taille minimale d'échantillon permettant d'atteindre un niveau de précision σ.

La formule (3.10) fait appel à a_h qui est l'allocation de l'échantillon dans la strate h. Plusieurs méthodes existent pour définir cette répartition, mais nous n'en retenons et comparons que deux : l'allocation proportionnelle et l'allocation x-optimale.

Allocation proportionnelle

Cette méthode consiste à répartir, comme son nom l'indique, de manière proportionnelle l'échantillon dans chacune des H^* strates (il a été décidé que la dernière strate serait entièrement télérelevée - au vu du faible effectif et de la forte disparité des comportements, cette contrainte semble justifiée). Ainsi la répartition de l'échantillon dans une strate dépend du poids de la strate dans la population :

$$a_h = \frac{n_h}{n^*} = \frac{N_h}{N^*} \Longleftrightarrow n_h = N_h \times \frac{n^*}{N^*} \qquad \forall h \in [\![1, H^*]\!]. \tag{3.11}$$

où $N^* = N - N_H$. Une conséquence pratique à l'utilisation d'une allocation proportionnelle est que le taux de sondage est identique d'une strate à l'autre :

$$f_h = \frac{n_h}{N_h} = \frac{n^*}{N^*} \qquad \forall h \in [\![1, H^*]\!]. \tag{3.12}$$

Allocation x-optimale

Cette méthode est dérivée de l'allocation de Neyman qui, en plus de prendre en compte l'effectif de chaque strate, inclut la dispersion de celle-ci. L'allocation de Neyman [Neyman, 1934] est la répartition de l'échantillon qui minimise la variance de l'estimateur et est solution du problème d'optimisation suivant :

$$\begin{aligned} \min_{n_1,\dots,n_{H^*}} \quad & \sum_{h=1}^{H^*} N_h^2 \frac{1-f_h}{n_h} S_{Yh}^2 \\ \text{s.c.} \quad & \sum_h n_h = n^*. \end{aligned}$$

La solution de ce problème renvoie :

$$n_{h-Neyman} = n^* \times \frac{N_h S_{Yh}}{\sum_{i=1}^{H^*} N_i S_{Yi}} \qquad \forall h \in [\![1, H^*]\!].$$

Cependant comme précédemment, les valeurs S_{Yh} sont inconnues. L'allocation x-optimale préconise alors d'utiliser la variable de stratification :

$$n_{h-x.opt} = n^* \times \frac{N_h S_{Xh}}{\sum_{i=1}^{H^*} N_i S_{Xi}} \qquad \forall h \in [\![1, H^*]\!]. \tag{3.13}$$

La mise en oeuvre de ces deux méthodes sur notre population renvoie des allocations différentes comme le montre la Figure 3.6

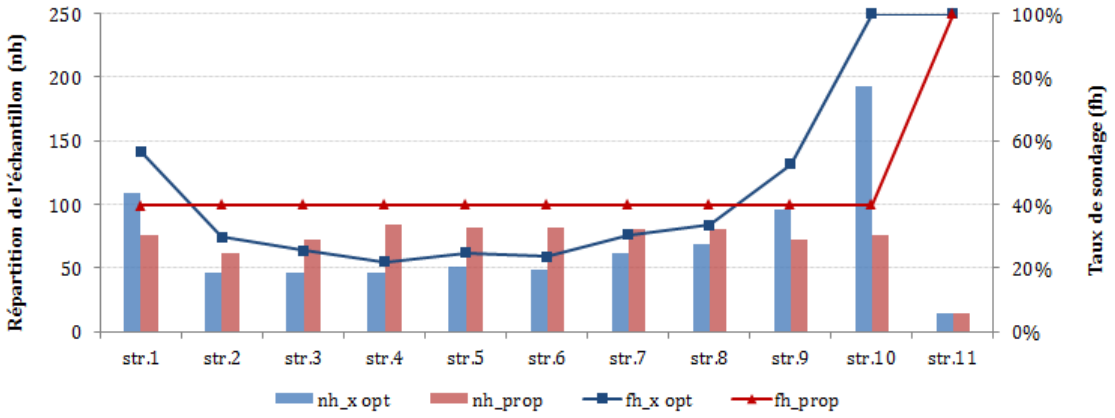


FIGURE 3.6 – Allocation de l'échantillon en fonction des 2 méthodes

L'allocation proportionnelle attribue un nombre quasi-identique d'individus sondés dans chaque strate, même si elles n'ont pas toutes la même dispersion (*cf.* Table 3.4).

La strate 4 fait partie des strates ayant la plus faible dispersion, pourtant elle est la strate la mieux représentée en termes de nombre d'individus. La strate 10, quant à elle, a une dispersion 1000 fois supérieure et pourtant elle a un effectif inférieur à la strate 4. Cet écart entre le nombre d'individus sondés et la dispersion au sein des strates montre l'inadéquation de la méthode de l'allocation proportionnelle à notre cas d'étude. C'est donc la méthode de l'allocation x-optimale qui est retenue.

TABLE 3.4 – Comparaison des méthodes de répartition de l'échantillon

strate	N_h	S_h^2	N_h/N	n_h**-x opt**	f_h**-x opt**	n_h**-prop**	f_h**-prop**
1	192	105	10%	109	57%	76	40%
2	155	28	8%	46	30%	61	40%
3	184	21	9%	47	26%	73	40%
4	213	16	11%	47	22%	84	40%
5	206	19	11%	51	25%	82	40%
6	207	18	11%	49	24%	82	40%
7	203	30	10%	62	31%	80	40%
8	205	36	10%	69	34%	81	40%
9	182	90	9%	96	53%	72	40%
10	193	20 611	10%	193	100%	77	40%
11	14	1 180 003	1%	14	100%	14	100%

Si nous revenons maintenant à l'égalité (3.10), en remplaçant a_h par la forme donnée à l'équation (3.13), il est possible de calculer la taille d'échantillon minimale nécessaire pour atteindre un écart-type de l'estimateur de consommation suffisamment faible pour permettre de détecter une fuite sur le réseau d'eau. On estime que le débit d'une fuite détectable sur un branchement d'eau potable est égal à 13 m^3/jour (ce qui est en adéquation avec les seuils des appareils de mesure de 0.4 m^3/h). Ainsi, nous cherchons à obtenir une estimation dont l'écart-type est inférieur ou égal à 13m^3. Il est important de noter que nous avons omis, par souci de simplicité, l'indice de temps t. Cependant dans la mesure où la variable Y dépend du temps, les quantités S_{Yh}^2 et $V(\hat{T}_Y)$ dépendent elles aussi du temps. De fait, d'après l'équation (3.10), la taille n de l'échantillon et sa répartition n_h au sein de la strate h doivent elles aussi être indicées par le temps. Il est alors possible de calculer $\mathscr{T}$ valeurs de la taille d'échantillonnage $n(t)$ solutions de l'équation (3.10), que l'on peut analyser avec le Tableau 3.5.

TABLE 3.5 – Statistiques sur les $\mathscr{T}$=365 tailles d'échantillon.

	1$^{\text{er}}$ Quart.	**Médiane**	**Moyenne**	**3$^{\text{ème}}$ Quart.**
n(t)	547	781	996	1 426
f(t)	28%	40%	51%	73%

Comme il a été indiqué précédemment, l'échantillon ne peut être modifié une fois créé, c'est pourquoi $n(t)$ doit être constant et fixé à une valeur n. Nous nous servons alors de la médiane (qui n'est pas influencée par les valeurs extrêmes) pour décider du nombre d'individus à échantillonner : on prend ainsi $n = 781$, ce qui correspond à un taux de sondage $f = 40\%$.

Les individus sont ensuite tirés par sondage aléatoire simple dans chaque strate suivant l'allocation x-optimale qui est préférable dans un cas comme le nôtre où il existe une forte

disparité des S_{Xh} (voir le plan de sondage du Tableau 3.6).

TABLE 3.6 – Définition du plan de sondage stratifié pour le calcul de l'estimateur

Strate	Bornes	N_h	S^2_{Xh}	N_h/N	n_h	n_h/N_h
1	[0 ; 30 [	192	105	10%	109	57%
2	[30 ; 49 [	155	28	8%	46	30%
3	[49 ; 65 [	184	21	9%	47	26%
4	[65 ; 79 [	213	16	11%	47	22%
5	[79 ; 94 [	206	19	11%	51	25%
6	[94 ; 109 [	207	18	11%	49	24%
7	[109 ; 129 [	203	30	10%	62	31%
8	[129 ; 150 [	205	36	10%	69	34%
9	[150 ; 185 [	182	90	9%	96	53%
10	[185 ; 1000 [	193	20 611	11%	193	100%
11	[1000 ; +∞ [	14	1 180 003	1%	14	100%

3.1.3 Estimation du total des consommations journalières

Notre échantillon s ayant été constitué suite à un sondage stratifié, il est possible maintenant de calculer l'estimateur du total des consommations journalières $\hat{T}_Y(t)$ (*cf.* équation (3.1)). Cet estimateur est un estimateur de Horvitz-Thompson, ce qui implique qu'il a pour propriété d'être sans biais : $\mathbb{E}[\hat{T}_Y(t)] = T_Y(t)$.

Pour pouvoir évaluer numériquement ce plan d'échantillonnage, nous avons opté pour une approche de type Monte-Carlo. A partir des données de la commune de Canéjan entièrement télérelevée, nous avons répliqué 25 000 fois le tirage d'un échantillon suivant ce plan. Les 25 000 simulations permettent de calculer un estimateur moyen journalier ainsi que son écart-type. Les résultats sont présentés partiellement à la Figure 3.7 et complètement en Annexe B.

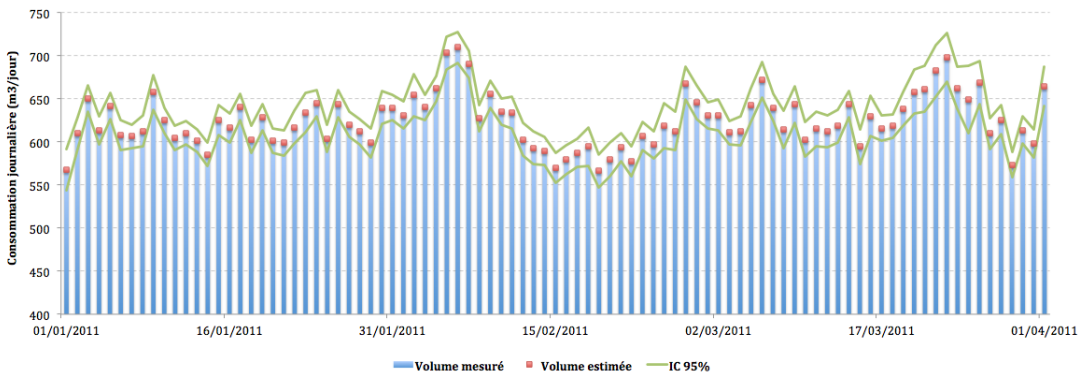


FIGURE 3.7 – Consommations journalières mesurées, estimées et intervalles de confiance à 95% pour les estimateurs du total de la consommation journalière

Les simulations illustrent clairement l'absence de biais de l'estimateur comme illustré à la Figure 3.7. Une majorité (52%) des écart-types sont inférieurs au seuil imposé (*cf.* Table 3.7), l'écart-type médian étant de 13 m^3/jour. Les résultats obtenus apparaissent donc satisfaisants.

TABLE 3.7 – Fractiles des estimations de l'écart-type des estimateurs de la consommation journalière sur 2011.

1^{er} Quartile des écart-types journaliers	10
Médiane des écart-types journaliers	13
Moyenne des écart-types journaliers	16
$3^{ème}$ Quartile des écart-types journaliers	18
% d'écart-types $\leq$ 13 m^3/j	52%

Le plan d'échantillonnage proposé dans cette première partie remplit convenablement les contraintes imposées : obtenir un estimateur du total des consommations qui a un écart-type inférieur ou égal à 13 m^3/jour. Cependant, il faut souligner le fait que la stratification proposée a été réalisée pour une estimation des consommations journalières de l'année 2011. Dans la mesure où nous avons affaire à des données longitudinales, dépendantes largement du comportement des individus, cet échantillon sera-t-il toujours adéquat pour une utilisation ultérieure, par exemple pour estimer les consommations 2012 ?

3.2 Les limites de l'estimateur par sondage stratifié

L'échantillon construit précédemment à l'aide de la consommation annuelle 2010 répond aux exigences imposées : les estimateurs de la consommation journalière totale en 2011 obtenus ont une précision médiane inférieure à 13 m^3. Le coefficient de corrélation entre la variable de stratification et la variable d'intérêt est d'autant plus fort que les deux variables sont proches temporellement, comme le montre la Figure 3.1. Autrement dit, la qualité de l'estimateur, fondé sur une stratification à l'année A, devrait avoir tendance à se dégrader au fur et à mesure de son utilisation au cours du temps. De plus, l'évolution des comportements des usagers entraîne des mouvements de strates (*stratum jumper*, [Rivest, 1999]) qui va dégrader l'homogénéité initiale des strates.

Considérons maintenant une nouvelle période d'étude, l'année 2012. Les consommations annuelles de 2011 sont donc disponibles, nous nous en servons pour redéfinir l'appartenance d'un individu à une strate telle que définie à la Table 3.6 selon sa consommation annuelle 2011. La Figure 3.8 met bien en évidence le phénomène de stratum jumper et l'évolution des strates au cours du temps.

On peut remarquer par exemple que seulement 37% des individus qui étaient en strate 8 d'après leur consommation annuelle en 2010 le restent en 2011. L'évolution de la population et des comportements individuels ne garantit plus l'homogénéité des strates, ce

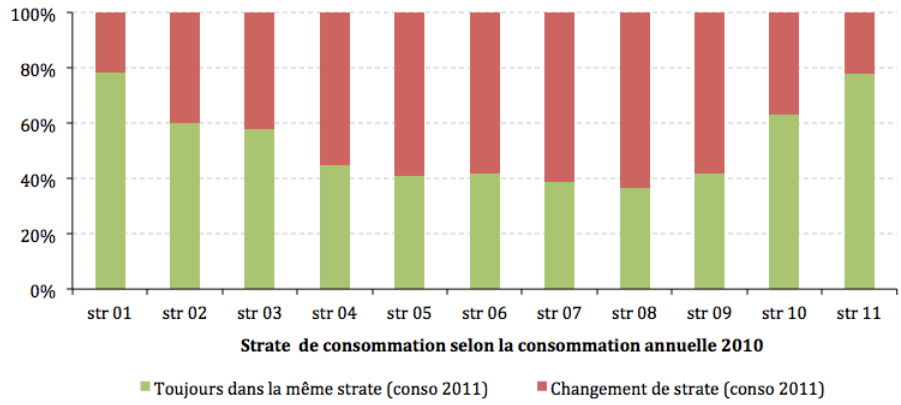


FIGURE 3.8 – Évolution de la composition des strates entre 2010 et 2011

qui dégrade la qualité de notre estimateur stratifié. L'estimation des consommations journalières de 2012 suivant le plan de sondage proposé en première partie, ne répond plus aux contraintes imposées. L'estimateur reste quoi qu'il en soit un estimateur du total sans biais comme on peut le voir sur la Figure 3.9 qui confronte les valeurs de consommation mesurée à la moyenne d'estimations obtenues par Monte Carlo.

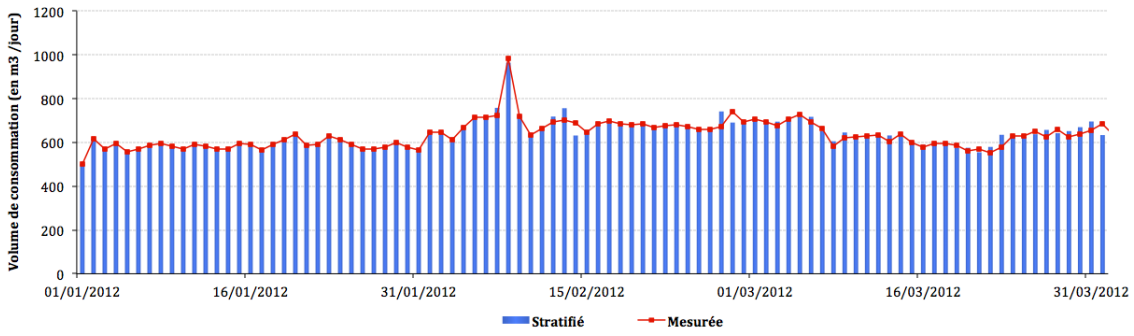


FIGURE 3.9 – Estimation de la consommation journalière totale par sondage stratifié sur la période du 01/01/2012 - 01/04/2012

La précision des estimateurs, voir à la Table 3.8, s'est quant à elle dégradée avec le temps :

TABLE 3.8 – Fractiles des estimations de l'écart-type des estimateurs du total de la consommation journalière en 2012 par sondage stratifié

1^{er} Quartile des écart-types journaliers	12
Médiane des écart-types journaliers	15
Moyenne des écart-types journaliers	20
$3^{ème}$ Quartile des écart-types journaliers	24
% d'écart-types $\leq$ 13 m^3/j	39%

L'utilisation de notre échantillon au cours du temps a pour conséquence de dégrader la qualité de l'estimateur : seuls 39% des écarts-types journaliers répondent à la contrainte imposée contre 52% auparavant.

Cependant, dans la mesure où nous cherchons à estimer les consommations journalières de l'année 2012, les consommations annuelles de l'année 2011 sont donc disponibles. Il convient de prendre en compte cette nouvelle information auxiliaire actualisée, qui est mieux corrélée à notre variable d'intérêt.
Afin de pallier la dégradation potentielle de la qualité de l'estimateur stratifié, nous allons procéder à un redressement de cet estimateur. Le redressement est une technique qui consiste à corriger l'estimateur grâce à une information auxiliaire corrélée à la variable d'intérêt (dans notre cas, la variable de redressement sera naturellement la consommation annuelle 2011). La variable utilisée pour le redressement sera notée Z, la variable d'intérêt (consommation journalière) est toujours notée Y.

3.3 Redressement de l'estimateur par sondage stratifié

Une erreur naturelle serait de reformer les strates avec cette information auxiliaire afin de reformer des groupes homogènes. Dans ce cas, il ne s'agira pas d'un échantillon issu d'un sondage stratifié car le tirage des individus n'est plus aléatoire : il est conditionné par le tirage initial. Il faut "post-stratifier" notre estimateur, c'est-à-dire recréer des strates adaptées à la nouvelle variable auxiliaire, tout en prenant en compte dans le calcul de l'estimateur des probabilités d'inclusion initiale.

3.3.1 Redressement par post-stratification

Cette méthode permet de stratifier non pas la population, mais l'échantillon déjà extrait, selon la variable auxiliaire actualisée. A priori, cette opération est appropriée compte tenu de la variable d'intérêt considérée. « *Toute opération de post-stratification consécutive à un sondage aléatoire simple [...] améliore l'estimation par rapport à la moyenne simple* » [Ardilly, 2006]. Nous illustrons dans cette section l'effet de la post-stratification sur l'estimateur stratifié. La méthodologie développée en section 1 est ré-appliquée afin de découper la population selon la variable Z. On note :

- Γ_k, $k = 1, \ldots, K$, les post-strates de tailles respectives M_k,
- $A_{kh} = \Gamma_k \bigcap G_h$, l'intersection entre la strate G_h et la post-strate Γ_k,
- Θ_{kh}, l'effectif de A_{kh},
- γ_k l'intersection de l'échantillon s et de la post-strate Γ_k,
- m_k la taille aléatoire de l'échantillon γ_k,
- $\alpha_{kh} = \gamma_k \bigcap g_h$ l'intersection entre le sous-échantillon g_h et la post-strate Γ_k,
- θ_{kh} la taille aléatoire de l'échantillon α_{kh}.

.

L'estimateur post-stratifié du total s'écrit de la façon suivante :

$$\hat{T}_{Y\,post} = \sum_{k=1}^{K} M_k \frac{\sum_{i \in \gamma_k} d_i Y_i}{\sum_{i \in \gamma_k} d_i} \tag{3.14}$$

où $d_i = \frac{1}{\pi_i}$. Ainsi, deux estimateurs redressés par post-stratification sont envisageables, en fonction de l'estimateur initial considéré.

Approche 1 : redressement de l'estimateur du total stratifié

Comme nous l'avons vu dans l'équation (3.1), les poids de sondage π_i sont égaux à $\frac{n_h}{N_h}\mathbb{1}_{[i\in G_h]}$. En les ré-injectant dans l'équation (3.14), nous obtenons l'estimateur redressé suivant :

$$\begin{aligned}
\hat{T}_{Y\,post\,1} &= \sum_{k=1}^{K} M_k \frac{\sum_{i\in\gamma_k} \frac{N_h}{n_h} Y_i \mathbb{1}_{[i\in G_h]}}{\sum_{i\in\gamma_k} \frac{N_h}{n_h} \mathbb{1}_{[i\in G_h]}}, \\
&= \sum_{k=1}^{K} M_k \frac{\sum_{h=1}^{H} \sum_{i\in\alpha_{hk}} \frac{N_h}{n_h} Y_i}{\sum_{h=1}^{H} \sum_{i\in\alpha_{hk}} \frac{N_h}{n_h}}, \\
&= \sum_{k=1}^{K} M_k \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i\in\alpha_{hk}} Y_i}{\sum_{h=1}^{H} \frac{N_h}{n_h} \theta_{hk}}.
\end{aligned} \tag{3.15}$$

Approche 2 : redressement de l'estimateur du sous-total par strate

L'estimateur stratifié du total est en réalité une somme d'estimateurs par sondage aléatoire simple du total. Au lieu de redresser l'estimateur du total stratifié, nous allons redresser les H estimateurs des totaux par strate. On ne s'intéresse plus ici aux post-strates mais aux intersections strates/post-strates. Les poids de sondage initiaux π_i valent $\frac{n_h}{N_h}$, et l'estimateur redressé du total dans la strate h s'écrit :

$$\begin{aligned}
\hat{T}_{Y\,post\,h} &= \sum_{k=1}^{K} \Theta_{hk} \frac{\sum_{i\in\alpha_{hk}} \frac{N_h}{n_h} Y_i}{\sum_{i\in\alpha_{hk}} \frac{N_h}{n_h}}, \\
&= \sum_{k=1}^{K} \Theta_{hk} \frac{\frac{N_h}{n_h} \sum_{i\in\alpha_{hk}} Y_i}{\frac{N_h}{n_h} \sum_{i=1}^{\Theta_{hk}} \mathbb{1}_{[i\in\alpha_{hk}]}}, \\
&= \sum_{k=1}^{K} \frac{\Theta_{hk}}{\theta_{hk}} \sum_{i\in\alpha_{hk}} Y_i.
\end{aligned}$$

En sommant le total redressé par strate sur chacune d'entre elles, nous obtenons l'estimateur du total redressé par post-stratification :

$$\hat{T}_{Y\,post} = \sum_{h=1}^{H} \sum_{k=1}^{K} \frac{\Theta_{hk}}{\theta_{hk}} \sum_{i\in\alpha_{hk}} Y_i. \tag{3.16}$$

Nous pouvons remarquer que l'équation (3.16) impose que tous les sous-échantillons α_{hk} ne soient pas nuls pour que l'estimateur post-stratifié selon l'approche 2 soit calculable. A l'inverse, dans le cadre de l'approche 1, si au moins un sous-échantillon α_{hk} n'est pas vide, alors l'estimateur post-stratifié est calculable. On peut aussi noter que si $\Theta_{hk} = 0$ (l'intersection au sein de la population est vide) alors le problème est facilement soluble

en imposant $\frac{\Theta_{hk}}{\theta_{hk}} = 0, \forall \Theta_{hk} = 0$.

Appliquons ces deux méthodes de redressement par post-stratification sur les estimateurs de la consommation journalière pour l'année 2012. Au préalable, il convient de définir les post-strates Γ_k, c'est-à-dire leur nombre K ainsi que leurs bornes. Pour cela nous ré-appliquons la méthode vue auparavant pour définir les post-strates selon la consommation annuelle 2011 . Ainsi il convient de créer 18 post-strates (*cf.* Table 3.9). Comme il a été dit précédemment, le risque majeur de superposer une post-stratification sur une stratification est de se retrouver avec des intersections vides.

TABLE 3.9 – Post-strates définies sur la population

Post-str	M_k	M_k/N	Borne inf	**Post-str**	M_k	M_k/N	Borne inf
1	108	6%	0	**10**	116	6%	102
2	98	5%	21	**11**	114	6%	111
3	110	6%	36	**12**	106	6%	122
4	109	6%	48	**13**	104	6%	133
5	121	7%	58	**14**	98	5%	146
6	112	6%	67	**15**	92	5%	162
7	114	6%	75	**16**	89	5%	187
8	114	6%	84	**17**	85	5%	231
9	120	7%	93	**18**	12	1%	1000

L'application numérique du redressement nous a permis de tirer les résultats suivants : sur nos 25 000 simulations, toutes se sont retrouvées dans un cas où $\theta_{hk} = 0$ (pour $\Theta_{hk} \neq 0$). En moyenne dans une simulation, 25% des θ_{hk} sont nuls. On peut aussi analyser le problème dans l'autre sens (Table 3.10) : pour chaque intersection nous calculons le pourcentage de simulations (parmi les 25 000) où l'échantillon est vide.

On remarque que sur certaines intersections, en majorité les intersections sont vides : dans 75% des simulations, l'intersection de la strate 2 et de la post-strate 15 est vide. Ainsi la problématique d'échantillon nul au sein des intersections est toujours au moins une fois présente dans les simulations, ce qui empêche de calculer l'estimateur redressé de l'équation 3.16.

Des méthodes permettent cependant de résoudre ce problème en regroupant des "strates" adjacentes. [Jay et al., 2007] préconisent par exemple de regrouper les intersections pour éviter d'avoir des sous-échantillons vides. Supposons qu'il existe deux entiers h et k tels que $\alpha_{hk} = 0\,(A_{hk} \neq 0)$, la fusion des intersections peut se faire de la façon suivante :

$$A_{\{h,i\}k} = A_{hk} \bigcup A_{ik} \qquad \text{ou} \qquad A_{h\{k,j\}} = A_{hk} \bigcup A_{hj} \qquad (i \neq h, j \neq k)$$

tel que $\theta_{\{h,i\}k} \neq 0$ ou $\theta_{h\{k,j\}} \neq 0$.

TABLE 3.10 – Pourcentage de simulations pour lesquelles θ_{hk} est nul

		Strates										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Post-strates	**1**	0%	17%	78%	×	×	×	×	×	×	×	×
	2	0%	0%	29%	57%	44%	34%	44%	×	×	×	×
	3	0%	0%	0%	5%	26%	70%	67%	0%	0%	×	×
	4	1%	0%	0%	0%	58%	34%	×	0%	×	×	×
	5	×	13%	0%	0%	15%	8%	45%	0%	0%	×	×
	6	1%	47%	1%	0%	1%	8%	44%	0%	0%	×	×
	7	×	78%	29%	0%	0%	16%	9%	×	0%	75%	×
	8	43%	37%	48%	1%	0%	0%	1%	0%	0%	74%	×
	9	1%	61%	48%	45%	0%	0%	0%	0%	0%	×	×
	10	43%	×	37%	47%	3%	0%	0%	0%	0%	50%	×
	11	43%	56%	×	43%	9%	0%	0%	0%	0%	12%	×
	12	×	×	×	32%	77%	0%	0%	0%	0%	50%	×
	13	×	×	×	75%	20%	23%	0%	0%	0%	6%	×
	14	×	×	×	76%	45%	8%	4%	0%	0%	1%	×
	15	×	75%	78%	×	34%	70%	44%	0%	0%	0%	×
	16	43%	×	78%	×	59%	×	20%	0%	0%	0%	×
	17	×	×	×	75%	×	×	×	0%	0%	0%	56%
	18	×	×	×	×	×	×	×	×	×	56%	3%

× signifie que $\Theta_{hk} = 0$

Par exemple, dans notre application, l'intersection de la strate 1 et de la post-strate 8 a un effectif de 1 individu ($\Theta_{18} = 1$) alors que α_{18} est vide ($\theta_{18} = 0$). L'intersection adjacente (A_{19}) est composée de 2 individus dont un faisant partie de l'échantillon ($\Theta_{19} = 2$ et $\theta_{19} = 1$). Ces deux groupes sont alors fusionnés pour former un groupe composé de 3 individus ($\Theta_{1\{8,9\}} = 3$) dont 1 a été échantillonné ($\theta_{1\{8,9\}} = 1$).

Ainsi il est possible de calculer les deux estimateurs redressés par post-stratification. Sur l'intégralité de la période 2012, nous calculons dans un premier temps l'estimateur stratifié du total des consommations journalières que nous redressons ensuite par post-stratification. Les résultats des simulations d'échantillonnage et de redressement sont présentés à la Figure 3.10 et la Table 3.11. Pour une meilleure lisibilité des résultats, les graphiques n'illustrent qu'une partie des résultats, mais les tableaux, quant à eux, résument l'intégralité des résultats obtenus à partir des 366 estimations de la consommation journalière totale en 2012.

La Figure 3.10 illustre parfaitement le fait que les estimateurs redressés par post-stratification sont toujours des estimateurs sans biais.

Cependant, on remarque que les résultats de précision de ces estimateurs redressés (*c.f.* Table 3.11) ne sont pas forcément concluants (en particulier la post-stratification des sous-totaux).

Comme attendu, la présence de sous-échantillons vides dans les intersections strates/post-

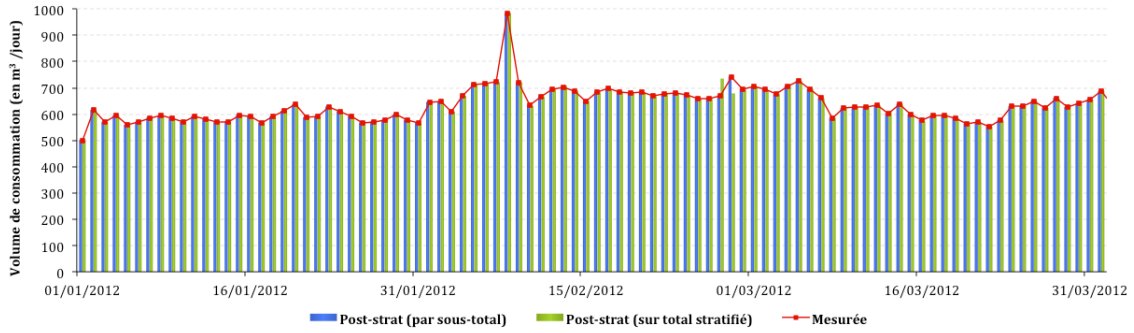


FIGURE 3.10 – Estimation de la consommation journalière totale par sondage post-stratifié sur la période du 01/01/2012 - 01/04/2012

TABLE 3.11 – Fractiles des estimations de l'écart-type des estimateurs du total de la consommation journalière en 2012 redressés par post-stratification

	Stratification initiale	Post-strat. (sur le total)	Post-strat. (sur les sous-totaux)
1er Quartile	12.2	12.0	23.7
Médiane	15.8	15.5	37.1
Moyenne	19.9	19.6	41.3
3ème Quartile	23.9	23.4	53.0
% Ecart-type $\leq 13 m^3/j$	34%	36%	2%

strates dégrade la précision de l'estimateur (une précision médiane de 15.8 m^3/jour pour l'estimateur stratifié contre 37.1 m^3/jour après post-stratification sur les sous-totaux). Même si la fusion des intersections semblait être une solution adéquate pour pallier ce problème, cela ne suffit pas à améliorer la précision de l'estimateur initial. Néanmoins un redressement sur le total semble être performant même si la réduction de l'écart-type est relativement faible (2%). Cette méthode, qui semblait naturelle après un sondage stratifié préalable, trouve ses limites dans cette même stratification.

Nous allons, dans la partie suivante, essayer une deuxième méthode de redressement : le redressement par régression. Comme il est judicieusement rappelé dans [Särndal et al., 1992], le redressement par post-stratification est un cas particulier de redressement par régression pour lequel les variables auxiliaires sont des variables indicatrices de l'appartenance à une strate.

3.3.2 Redressement par régression

Étant donné la forte corrélation linéaire entre la variable d'intérêt Y et la variable Z, il est raisonnable de supposer l'existence d'une relation de type affine entre ces deux variables :

$$Y_i(t) = \alpha(t) + \beta(t) Z_i + \varepsilon_i(t), \qquad \forall i \in [\![1, N]\!]$$

avec $\sum_{i\in U}\varepsilon_i(t)=0$.

Ainsi, une piste à exploiter est le redressement par régression (voir par exemple [Särndal et al., 1992]).

Notons $\widetilde{\mathbf{Z}}_i$ un vecteur de variables auxiliaires (défini par la suite). Nous considérons $\hat{T}_{\tilde{Z}}$ l'estimateur du total $T_{\tilde{Z}}$. Le redressement par régression de $\hat{T}_Y(t)$ s'effectue de la façon suivante :

$$\hat{T}_{Yr}(t)=\hat{T}_Y(t)+\hat{\mathbf{B}}'(t)\left(\mathbf{T}_{\widetilde{\mathbf{Z}}}-\hat{\mathbf{T}}_{\widetilde{\mathbf{Z}}}\right)=\hat{T}_Y(t)+\hat{\mathbf{B}}'(t)\left(\mathbf{T}_{\widetilde{\mathbf{Z}}}-\sum_{i\in s}d_i\widetilde{\mathbf{Z}}_i\right), \tag{3.17}$$

où

$$\hat{\mathbf{B}}(t)=\left(\sum_{i\in s}d_i\widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i'\right)^{-1}\left(\sum_{i\in s}d_i\widetilde{\mathbf{Z}}_iY_i(t)\right). \tag{3.18}$$

Deux approches sont possibles, comme dans le cas de la post-stratification : soit on redresse l'estimateur sur toute la population, soit sur chacune des strates.

Approche 1 : redressement de l'estimateur du total

La première option consiste à redresser, suivant la formule (3.17), l'estimateur du total $\hat{T}_Y(t)$. Dans ce cas, $\widetilde{\mathbf{Z}}_i'=(1,Z_i)$ et $\hat{\mathbf{B}}(t)=(\hat{\alpha}(t),\hat{\beta}(t))'$. Le paramètre de pente $\hat{\beta}(t)$ est un scalaire, commun à toute la population :

$$\hat{\beta}(t)=\frac{\sum_{i\in s}d_i(Z_i-\frac{\hat{T}_{Z\pi}}{N})(Y_i(t)-\frac{\hat{T}_{Y\pi}(t)}{N})}{\sum_{i\in s}d_i(Z_i-\frac{\hat{T}_{Z\pi}}{N})^2}\qquad\text{et}\qquad d_i=\frac{1}{\pi_i}=\frac{N_h}{n_h}\mathbb{1}_{[i\in G_h]}.$$

L'estimateur redressé est alors noté $\hat{T}_{Yr_1}(t)$.

Approche 2 : redressement de l'estimateur des sous-totaux

Dans la seconde option, nous considérons l'estimateur du total comme une somme pondérée d'estimateurs des moyennes par strate (formule (3.1)). Ainsi, nous redressons par régression tous les estimateurs des moyennes $\hat{\bar{y}}_h(t)$; $\widetilde{\mathbf{Z}}_i'=(I_{1i},I_{1i}Z_i,I_{2i},I_{2i}Z_i,\cdots,I_{Li},I_{Li}Z_i)$ où les I_h $(h=1,\cdots,L)$ sont des variables indicatrices indiquant l'appartenance aux strates. Dans ce cas, $\hat{\mathbf{B}}(t)=(\hat{\alpha}_1(t),\hat{\beta}_1(t),\cdots,\hat{\alpha}_L(t),\hat{\beta}_L(t))'$, chaque strate ayant alors sa propre pente $\hat{\beta}_h(t)$ définie par

$$\hat{\beta}_h(t)=\frac{\sum_{i\in g_h}(Z_i-\overline{Z}_h)(Y_i(t)-\overline{Y}_h(t))}{\sum_{i\in g_h}(Z_i-\overline{Z})^2},$$

L'estimateur correspondant est noté $\hat{T}_{Yr_2}(t)$.

Ainsi les deux estimateurs se distinguent par la variable auxiliaire exploitée. Un première comparaison entre les deux méthodes permet de constaster que la première approche ne nécessite d'estimer qu'un vecteur de paramètres $(\hat{\alpha}(t), \hat{\beta}(t))'$ alors que la seconde nécessite d'estimer un vecteur $(\hat{\alpha}_h(t), \hat{\beta}_h(t))'$ pour chacune des 11 strates. Cependant, il est préférable de juger des performances des deux méthodes selon leur capacité à réduire la variance de l'estimateur stratifié initial.

Comparaison des deux estimateurs redressés

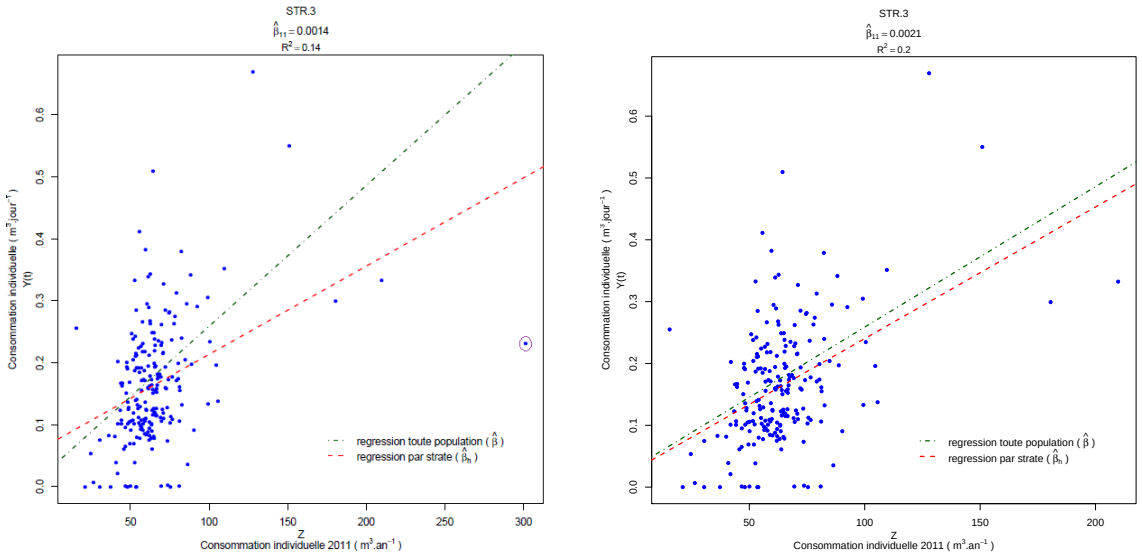
Le fait d'effectuer une régression par strate ne se justifie que dans le cas où les valeurs des coefficients de pente $\hat{\beta}_h(t)$ sont différents d'une strate à l'autre. Analysons le comportement entre la variable d'intérêt et la variable auxiliaire dans chaque strate. La variable d'intérêt étant la consommation individuelle journalière étudiée sur l'année 2012, il existe 366 régressions entre $Y(t)$ et Z. Nous ne présentons ici qu'un exemple (celui du 03/02/2012), mais les résultats numériques présentés ultérieurement en Table 3.13 traitent de l'ensemble des estimations journalières.

La pente estimée $\hat{\beta}(t)$ de la droite de régression calculée sur la population est égale à 2.3×10^{-3}. L'étude par strate, illustrée à la Figure C.1 présentée en Annexe C, suggère des relations linéaires entre la variable d'intérêt et la variable auxiliaire considérée. Les données de corrélations et de régression sont résumées à la Table 3.12.

TABLE 3.12 – Coefficients de régression estimés et coefficients de détermination

	$\beta(\mathbf{t})$	$\mathbf{R^2}$
Population	2.3×10^{-3}	0.56
strate 1	2.2×10^{-3}	0.54
strate 2	2.5×10^{-3}	0.12
strate 3	1.4×10^{-3}	0.38
strate 4	1.6×10^{-3}	0.28
strate 5	1.7×10^{-3}	0.26
strate 6	1.7×10^{-3}	0.17
strate 7	1.9×10^{-3}	0.38
strate 8	1.6×10^{-3}	0.29
strate 9	2.0×10^{-3}	0.47
strate 10	2.1×10^{-3}	0.24
strate 11	1.9×10^{-3}	0.66

L'étendue des $\hat{\beta}_h(t)$ va de 1.4×10^{-3} à 2.5×10^{-3}, pour une valeur moyenne de 1.9×10^{-3}, proche de la valeur de $\hat{\beta}$. L'ensemble des régressions linéaires sur les 11 strates de consommation est présenté à la Figure C.1 en Annexe C. Si on regarde en particulier la strate 3, dont le coefficient de pente est nettement différent de $\hat{\beta}$, on remarque (*cf.* Figure 3.11.a) que cette différence est due à la présence d'un seul point atypique (distance de Cook = 4.5). En effet, la droite calée sur la strate est très fortement influencée par ce point ayant une valeur importante pour Z (> 200 m^3/an). En filtrant la population sur cette strate pour en exclure cet individu, le coefficient de la pente de régression vaut 2.4×10^{-3}, comme on peut le voir à la Figure 3.11.b.



(a) (b)

FIGURE 3.11 – Régression linéaire entre $Y(t)$ et Z sur la strate 3 initiale (a) et filtrée (b)

La régression sur cette strate filtrée est donc très similaire à la régression sur toute la population, les droites de régression (sur la population et sur les strates) sont quasiment confondues. Par ailleurs, si on compare les coefficients de détermination (R^2), à une exception près (strate 11), la régression sur la population est mieux ajustée que celle sur les strates. A priori, la régression sur les sous-totaux ne semble pas se justifier dans notre cas, au vu des valeurs des $\hat{\beta}_h$ par rapport à $\hat{\beta}$ et des coefficients de détermination. Vérifions notre hypothèse sur l'ensemble de la période concernée en évaluant les estimateurs redressés.

L'application de ces deux méthodes de redressement à notre cas opérationnel permet de générer les résultats présentés ci-après. Comme pour la post-stratification, l'estimateur redressé par régression reste sans biais comme pour l'estimateur initial, comme on peut l'observer à la Figure 3.12.

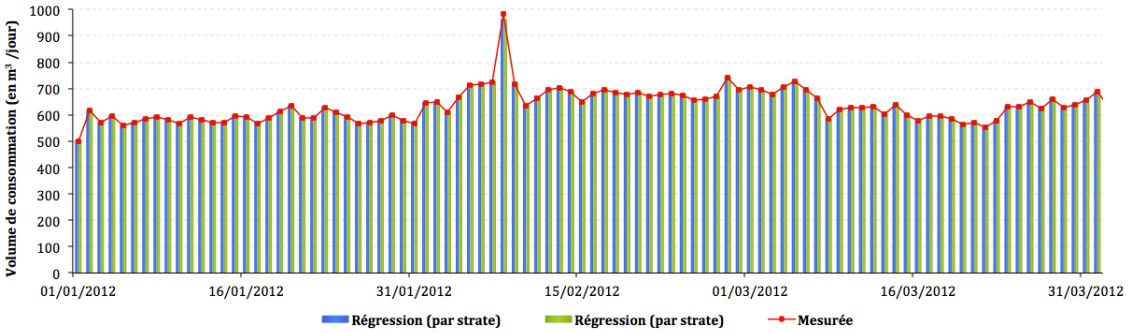


FIGURE 3.12 – Estimation de la consommation journalière totale par estimateur redressé par régression sur la période du 01/01/2012 - 01/04/2012

La réelle plus-value de cette méthode réside dans sa capacité à réduire l'écart-type

de l'estimateur initial. En effet comme l'attestent les résultats sur les 366 écart-types des estimateurs (stratifiés et redressés par régression) du total des consommations journalières (*cf.* Table 3.13), le redressement par régression, quelle que soit l'approche, a un effet mélioratif sur la précision des estimateurs.

TABLE 3.13 – Fractiles des estimations de l'écart-type des estimateurs du total de la consommation journalière en 2012 redressés par régression

	Stratification initiale	Redressement par rég. (sur le total)	Redressement par rég. (sur les sous-totaux)
1er Quartile	12.2	10.9	11.8
Médiane	15.8	14.0	15.4
Moyenne	19.9	18.2	19.5
3ème Quartile	23.9	21.9	24.0
% Ec.tp $\leq 13 m^3/j$	34%	46%	39%

Même si les deux méthodes semblent être avantageuses, le redressement par régression sur le total est la méthode la plus performante des deux, avec une réduction moyenne de l'écart-type de 10% ($\simeq$ - 2 m^3/jour) ainsi qu'une augmentation notable du nombre d'écarts-types journaliers répondant à la contrainte imposée.

3.3.3 Redressement par calage

Une troisième méthode de redressement est le calage. A chaque individu i de l'échantillon s est associé un poids de sondage d_i afin de construire l'estimateur de Horvitz-Thompson du total $\hat{T}_{Y\pi} = \sum_s d_i Y_i(t)$. L'idée du calage est de calculer, à partir des poids d_i, de nouveaux poids w_i en tenant compte de la variable auxiliaire Z. Notez que la méthode nécessite un vecteur de variables auxiliaires (comme décrit à la section suivante). Plus précisément, l'objectif est de trouver les poids w_i qui soient solutions du problème d'optimisation suivant :

$$\begin{cases} \min \sum_{i\in s} D(w_i, d_i) \\ \text{s.c.} \sum_{i\in s} w_i Z_i = \sum_{i\in U} Z_i, \end{cases}$$

où D est une mesure de distance. Notons que par abus de langage, nous citerons D comme une distance même si elle ne vérifie pas les 3 critères de définition d'une distance au sens mathématique du terme, notamment le fait que $D(d_i, w_i) = D(w_i, d_i)$. Les poids w_i doivent être proches des d_i, l'usage des poids initiaux garantissant un estimateur sans biais.

Si on note $\delta(w_i, d_i) = \partial D(w_i, d_i)/\partial w_i$, il est alors possible de définir une fonction F, nommée fonction de calage, telle que $d_i F(.)$ soit l'inverse de la fonction $\delta(., d_i)$ [Tillé, 2001], vérifiant

$$w_i = d_i F(\lambda Z_i), \tag{3.19}$$

où λ est le multiplicateur de Lagrange solution de l'équation :

$$\sum_{i\in s} d_i F(\lambda Z_i) Z_i = \sum_{i\in U} Z_i.$$

L'estimateur redressé par calage généralisé s'écrit :

$$\hat{T}_{Y\,Cal}(t) = \sum_{i\in s} w_i Y_i(t) = \sum_{i\in s} d_i F(\lambda Z_i) Y_i(t). \tag{3.20}$$

Deville et Särndal [Deville and Särndal, 1992] définissent une forme généralisée de la fonction de calage F à partir d'un réel α, forme depuis laquelle il est possible de calculer la mesure de distance D associée (***cf.*** Table 3.14).

TABLE 3.14 – Distance et fonction de calage définies selon [Deville and Särndal, 1992].

α	$D^\alpha(w_i,d_i)$	$F^\alpha(t)$
$\mathbb{R}\backslash\{0,1\}$	$\dfrac{\frac{w_i^\alpha}{d_i^{\alpha-1}} + (\alpha-1)d_i - \alpha w_i}{\alpha(\alpha-1)}$	$\sqrt[\alpha-1]{1+t(\alpha-1)}$
0	$-d_i \ln(\frac{w_i}{d_i}) + w_i - d_i$	
1	$w_i \ln(\frac{w_i}{d_i}) + d_i - w_i$	$\exp(t)$

Notons que pour le cas $\alpha = 1$, les nouveaux poids w_i sont toujours positifs, dans la mesure où les poids initiaux d_i le sont aussi (sondage stratifié). Le fait que ces nouveaux poids soient supérieurs ou inférieurs aux poids initiaux ne dépend que du signe du facteur λ (indépendamment de la valeur de la variable auxiliaire Z qui est toujours positive dans notre cas).

Il y a une infinité de possibilités de redressement par calage en fonction du choix de $\alpha \in \mathbb{R}$ intervenant dans la distance D^α, en plus des autres distances existantes (comme la distance euclidienne par exemple). Cependant nous n'en comparons que deux dans ce travail, les deux les plus couramment utilisées en pratique :

– le cas où $\alpha = 1$, cette méthode est appellée le *Raking Ratio*,
– le cas où $\alpha = 2$ (ici D^α est la distance du χ^2).

Les distances correspondantes sont :

$$D^1(d_i,w_i) = w_i \ln(\frac{w_i}{d_i}) + d_i - w_i \qquad \text{et} \qquad D^2(d_i,w_i) = \frac{(w_i-d_i)^2}{2d_i},$$

et les fonctions de calages associées sont les suivantes :

$$F^1(t) = \exp(t) \qquad \text{et} \qquad F^2(t) = 1+t.$$

Un des avantages du choix $\alpha = 1$ ou $\alpha = 2$ est que le problème d'optimisation mène toujours à une solution. Chacune des deux méthodes re-définit les poids initiaux d_i en :

$$\begin{aligned} w_i &= d_i F^1(\lambda Z_i) = d_i e^{\lambda Z_i} \quad \text{si } \alpha = 1, \\ w_i &= d_i F^2(\lambda Z_i) = d_i(1+\lambda Z_i) \quad \text{si } \alpha = 2. \end{aligned}$$

L'application de ces deux méthodes de redressement par calage ($\alpha = 1$ et $\alpha = 2$) avec pour variable auxiliaire Z les consommations annuelles individuelles de l'année 2011 a permis de générer les résultats présentés ci-après.

Tout comme pour les redressements par post-stratification et par régression, il n'y a aucune dégradation du biais de l'estimateur stratifié initial comme illustré à la Figure 3.13.

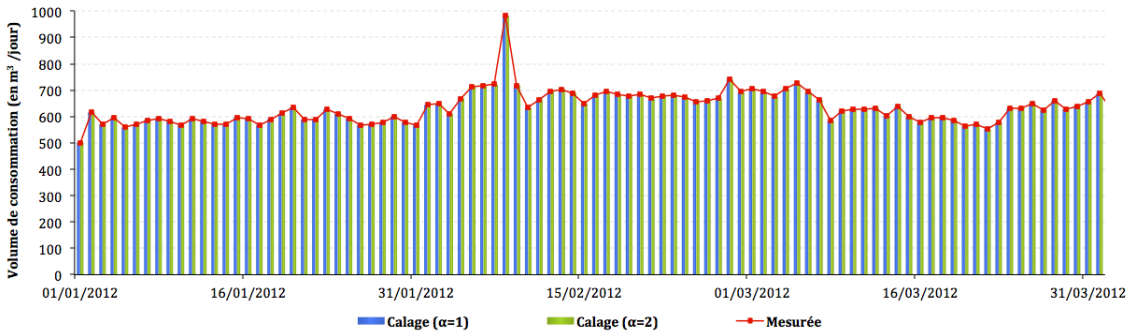


FIGURE 3.13 – Estimation de la consommation journalière totale par estimateur redressé par calage sur la période du 01/01/2012 - 01/04/2012

L'intérêt du redressement par calage est le fait de pouvoir réduire l'écart-type de l'estimateur stratifié initial. La précision des 366 estimateurs du total redressé est résumée dans la Table 3.15.

TABLE 3.15 – Fractiles des estimations de l'écart-type des estimateurs du total de la consommation journalière en 2012 redressés par calage

	Stratification initial	Calage ($\alpha = 1$)	Calage ($\alpha = 2$)
1er Quartile	12.2	11.9	11.9
Médiane	15.8	15.3	15.3
Moyenne	19.9	19.4	19.4
3ème Quartile	23.9	23.3	23.3
% Ec.tp $\leq 13 m^3/j$	34%	40%	40%

Les résultats en termes de précision sont équivalents pour les deux méthodes, ce qui confirme que les méthodes de calage, suivant une fonction F comme définie à la Table 3.14, sont asymptotiquement équivalentes ([Deville and Särndal, 1992]). Même si

les résultats d'estimation sont satisfaisants, le gain en précision est toutefois pauvre (réduction de la moyenne des écarts-types de 3%).

Remarque : Le redressement par régression (avec une seule variable auxiliaire Z) correspond à un cas particulier du redressement par calage : le calage avec une distance du χ^2 et le couple de variables auxiliaires (Z,1). La démonstration de cette relation est développée dans [Deville et al., 1993] ou [Ardilly, 2006].

3.3.4 Choix de la méthode de redressement

TABLE 3.16 – Synthèse des précisions obtenues par redressement.

	Strat.	Post-str. (tot.)	Post-str. (sous-tot.)	Régr. (tot.)	Régr. (sous-tot.)	Calage ($\alpha = 1$)	Calage ($\alpha = 2$)
1er Quartile	12.2	12.0	23.7	10.9	11.8	11.9	11.9
Médiane	15.8	15.5	37.1	14.0	15.4	15.3	15.3
Moyenne	19.9	19.6	41.3	18.2	19.5	19.4	19.4
3ème Quartile	23.9	23.4	53.0	21.9	24.0	23.3	23.3
% Ec.tp $\leq 13 m^3/j$	34%	36%	2%	46%	39%	39%	39%

Au vu des résultats du Tableau 3.16, on constate que le redressement par régression (sur le total) renvoie les meilleurs résultats en termes de réduction de l'écart-type de l'estimateur. Comme le montre la Figure 3.14, on remarque que l'écart-type de l'estimateur redressé par régression est uniformément inférieur à celui de l'estimateur stratifié.

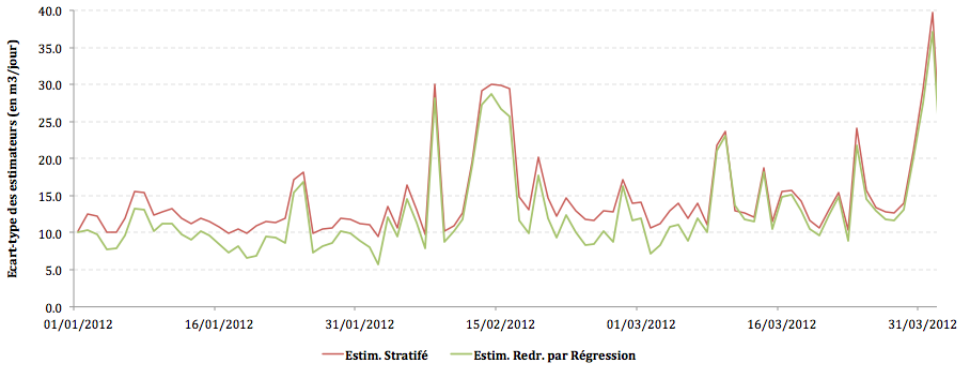


FIGURE 3.14 – Écarts-types journaliers des estimateurs issus d'un sondage stratifié et redressés par régression

Ainsi, le redressement est un opération efficace pour améliorer la précision des estimateurs. Même si le gain semble marginal, il est important de souligner que l'écart temporel qui sépare la création de l'échantillon de son utilisation n'est que d'un an. Plus cet écart grandit, plus les performances du redressement sur la réduction de la variance des estimateurs devraient être importantes.

3.4 Validation de la performance de la méthode d'équipement ciblé des compteurs

La validation de la méthode développée dans ce chapitre s'effectue en deux temps. Tout d'abord, nous jugeons de l'efficacité de construire un plan de sondage complexe (sur le nombre et les bornes des strates notamment), comparé à d'autres plans qui pourraient être qualifiés de plus naturels ou même plus simples. Puis nous jugeons la performance de notre plan de sondage en l'appliquant tel quel sur une nouvelle population.

3.4.1 Comparaison des méthodes d'échantillonnage

Nous avons présenté ici une méthodologie d'estimation de la consommation adapté à notre contrainte opérationnelle. Il existe d'autres méthodes d'échantillonnage ainsi que d'autres manières de construire un plan stratifié. Nous allons maintenant comparer notre méthodologie à d'autres qui peuvent paraître plus naturelles ou encore plus simples à implémenter. Nous allons comparer sept méthodes d'échantillonnage [1]. Les échantillons ont été construits en 2011 grâce à la consommation annuelle de 2010 et utilisés pour estimer la consommation journalière totale de 2012. Les sept méthodes sont les suivantes :

1 Sondage aléatoire simple (SAS), poids de sondage égaux pour tous les individus,
2 SAS, poids de sondage proportionnel à la consommation annuelle de 2010,
3 Sondage stratifié (STR), strate définies selon expertise métier, allocation x-optimale,
4 STR, strates définies selon expertise métier, allocation dans chaque strate proportionnelle aux consommations annuelles de 2010,
5 STR, strates définies selon expertise métier, allocation constante dans chaque strate,
6 Méthodologie présentée en section 3.1, sans redressement,
REF Méthodologie complète (avec redressement).

Les méthodes 3 à 5 font appel à une stratification définie selon une expertise métier ; il y a au total 8 strates définies à la Table 3.17.

TABLE 3.17 – Strates de consommation définies selon une expertise métier

Strate	1	2	3	4	5	6	7	8
Borne inférieure (m^3/an)	0	50	100	150	200	300	500	$\geq 1\,000$
Borne supérieure(m^3/an)	49 [2]	99	149	199	299	499	999	

Les sept méthodes renvoient toutes un estimateur sans biais ; seule la variance de l'estimateur peut varier. C'est sur ce critère que se base notre comparaison. Comme nous l'avons montré en section 1.2, la précision de l'estimateur est liée à la taille de l'échantillon. Donc partant d'une précision fixée, nous évaluons, pour nos sept méthodes, la taille d'échantillon nécessaire pour l'atteindre. Plus précisément, nous calculons la précision de l'estimateur obtenu grâce à notre méthodologie **REF**, qui nous servira de référence pour

1. Le plan de sondage ne concerne que les ménages, les *gros consommateurs* étant enquêtés exhaustivement.

2. Borne définie sur la base que la consommation moyenne maximale d'un foyer d'une personne serait de 50 m^3/an.

la comparaison. Le tableau 3.18 fournit, pour chaque méthode, le taux de sondage nécessaire pour atteindre la même précision que pour la méthode de référence.

TABLE 3.18 – Comparaison des différents plans de sondage

Méthode	Taille n d'échantillon	Taux f de sondage	Investissement suppl. (k€)
1	1 817	93%	73 (+132%)
2	1 934	99%	81 (+148%)
3	840	43%	4 (+8%)
4	899	46%	8 (+15%)
5	1 387	71%	42 (+78%)
6	821	42%	3 (+5%)
REF	781	40%	-

Pour chacune des six premières méthodes, nous avons calculé la taille d'échantillon nécessaire pour atteindre la même précision que celle obtenue avec la méthode de référence. Un premier coup d'œil au Tableau 3.18 nous permet de constater que les plans de sondage issus d'un sondage aléatoire simple (méthode 1 et 2) sont les méthodes qui requièrent le plus fort taux de sondage. En effet, l'écart-type d'un estimateur issu d'un sondage aléatoire simple est supérieur ou égal à celui d'un estimateur issu d'un sondage stratifié ([Cochran, 1977]). Ainsi, à précision égale, le sondage aléatoire requiert un plus fort taux de sondage qu'un sondage stratifié.

Ensuite si nous comparons les méthodes 3, 4, 5, on se rend compte de l'importance de l'allocation de l'échantillon dans chaque strate. En effet, vu la forte variabilité de la variable de stratification (un nombre important de faibles valeurs et un nombre restreint de fortes valeurs), l'allocation x-optimale est préférable dans notre cas d'étude.
Puis si nous comparons les méthodes 3 et 6, nous pouvons juger l'efficacité de notre stratification de la population qui semble fournir de meilleurs résultats.

Enfin la comparaison des méthodes 6 et REF montre l'importance de redresser un estimateur quand une nouvelle information utile est disponible.

La dernière colonne du Tableau 3.18 indique les investissement supplémentaires nécessaires [3] pour agrandir l'échantillon (partant d'un taux de sondage initial de 40%), sur la commune de Canéjan. L'utilisation de la méthode proposée, par rapport à des techniques de sondage plus simples, permet de réaliser des économies pouvant représenter jusqu'à 130% des investissements initiaux. Le redressement de l'estimateur à lui seul permet d'économiser 5% de dépenses en instrumentation. L'écart entre la stratification et le redressement n'étant que d'une année, les économies liées au redressement auront donc tendance à croître si cet écart s'agrandit.

Le redressement permet donc de réaliser des économies de budget, dans la mesure où cette technique ne nécessite aucun investissement, les données requises (base Clientèle) étant à la disposition de l'exploitant.

3. sous l'hypothèse que la pose d'un émetteur coûte 70€ (pose + main d'œuvre)

3.4.2 Validation de la méthode d'échantillonnage sur un autre site télérelevé

La méthodologie d'estimation qui a été construite tout au long de ce chapitre a permis de définir la composition d'un échantillon d'après une contrainte requise sur la précision des estimateurs. Si définir la composition des strates est un protocole facilement extrapolable sur d'autres cas d'étude, déterminer la taille optimale d'échantillon reste une tâche compliquée. Comme il a été dit précédemment, même si déterminer la taille d'échantillon est une étape difficile, il est parfois possible de l'approcher à partir de populations similaires [Kpedekpo, 1973]. Nous tentons alors de valider cette hypothèse en vérifiant si la commune de Canéjan pourrait servir de référence pour des populations identiques. En d'autres termes, la Figure 3.5, obtenue à partir des données de Canéjan, pourrait-elle être directement ré-utilisée sur une autre commune semblable à celle de Canéjan ?

Pour des raisons de confidentialité, cette partie n'est pas disponible.
Pour plus d'informations, merci de contacter l'auteur :
karim.claudio@lyonnaise-des-eaux.fr

3.5 Conclusion du chapitre

Nous avons développé, tout au long de ce chapitre, une méthodologie d'instrumentation des compteurs et d'estimation des consommations journalières totales. La méthode a été développée de sorte que l'échantillon construit s'adapte d'une part à nos données (données individuelles comportementales longitudinales) mais aussi à notre problématique industrielle (aide à la détection des pertes sur les réseaux d'eau potable).

Nous obtenons ainsi, grâce à ce plan d'instrumentation stratégique des compteurs, une estimation de la consommation journalière d'un secteur hydraulique, respectant les contraintes sur la précision requise. Mais le plan de sondage, aussi efficace soit-il, devient rapidement obsolète au cours du temps à cause de la dégradation de la corrélation entre la variable d'intérêt et la variable de stratification. L'estimateur initial restant malgré tout sans biais, c'est essentiellement la précision de cet estimateur qui se dégrade et qu'il est nécessaire d'améliorer. Les techniques de redressement, présentées en seconde partie, permettent quasiment toutes de répondre à cette problématique. Il est possible alors, en actualisant "régulièrement" les données, d'estimer fiablement les consommations journalières totales d'un secteur hydraulique.

A partir des estimateurs de la consommation, mais aussi des données de volumes journaliers délivrés au réseau , il est possible d'établir un premier niveau de la balance de l'eau (*cf.* Table 1.1). Notons $E(t)$ et $C(t)$ respectivement les volumes délivrés à un réseau et les consommations totales au temps t, nous pouvons ainsi calculer les pertes en eau, $L(t)$, sur ce secteur :

$$L(t) = E(t) - C(t).$$

Malheureusement, ce volume de pertes ne suffit pas à lui seul à optimiser la gestion du réseau et la lutte contre les fuites. Comme il a été précédemment mentionné, toutes

les fuites ne sont pas repérables et donc réparables, en fonction de leur état (*cf.* Fig. 1.1). C'est pourquoi il est nécessaire de détailler cette première version de la balance de l'eau en une description des pertes en fonction des trois états de fuite. L'objectif de la prochaine partie de ce manuscrit sera de développer un modèle permettant de décomposer le débit de fuite.

Chapitre 4

Décomposition du débit de fuite

L'objectif principal de l'estimation des consommations est de pouvoir in fine conduire à une estimation des pertes en eau potable. La technologie permet aujourd'hui de connaître les volumes livrés au réseau (VLAR) sur un pas de temps journalier. En soustrayant les consommations totales journalières estimées aux VLAR, il est possible d'estimer les pertes journalières et ainsi construire une première version de la Balance de l'Eau (*cf.* Table 1.1).

4.1 Estimation des pertes sur le réseau d'eau potable

Sur le secteur que nous avons étudié (Canéjan), nous pouvons calculer les pertes à partir de l'estimation des consommations vue au chapitre précédent. Nous sommes partis d'une hypothèse forte qui est que la précision des pertes est égale à la précision de l'estimateur des consommations, supposant de fait que les erreurs de métrologie, dues aux appareils de mesure des VLAR et des compteurs télérelevés, sont nulles ou en tout cas négligeables face à l'écart-type de l'estimateur du total des consommations. Nous illustrons à la Figure 4.1 les VLAR, les pertes estimées ainsi qu'un intervalle de confiance à 95% de l'estimateur des pertes journalières.

Cet intervalle de confiance dépend de l'écart-type de l'estimateur des pertes. Ce dernier intègre différentes composantes :

- l'incertitude des appareils de mesure du débit entrant (environ 1 à 2% du volume entrant),
- l'incertitude des compteurs de vente d'eau (environ 3%, variable selon le débit, l'âge du compteur et le modèle),
- l'écart-type de l'estimateur des consommations (variable selon le taux d'instrumentation).

Au vu des différents ordres de grandeurs des sources d'imprécision des pertes estimées, nous considérons que les erreurs de mesure sont négligeables face à l'écart-type des estimateurs de la consommations journalières. On peut noter également que lors de la pose d'émetteurs télérelevés, les compteurs sont renouvelés en cas d'incompatibilité avec les émetteurs télérelevés. Ainsi on peut considérer que le sous-comptage (erreur de mesure des compteurs d'eau) est proche de 0.

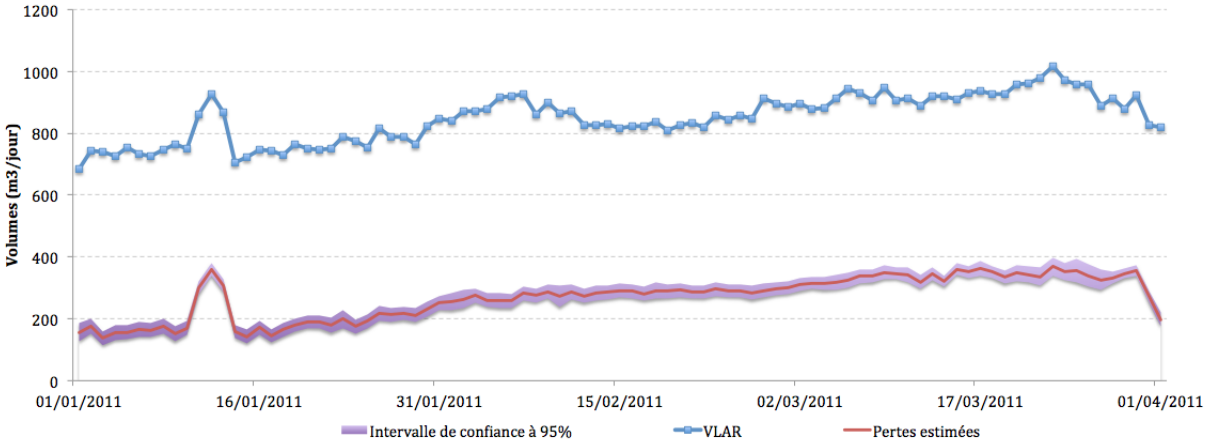


FIGURE 4.1 – Estimation des pertes en eau potable (01/01/2011 - 01/04/2011)

On peut attribuer une autre utilité à l'estimation des consommations grâce au télérelevé. En effet, les méthodes de détection des fuites comme la méthode du débit minimum nocturne (*cf.* chapitre 2.2.1) supposent une consommation nocturne constante, ce qui est rarement le cas si on considère un secteur hydraulique où se trouvent des industriels, sans compter les consommations domestiques nocturnes exceptionnelles. Sur un secteur comme celui de la Cote 50 à Cenon, où le télérelevé fournit des informations toutes les heures, il est possible de calculer ces consommations nocturnes, comme illustré à la Figure 4.2.

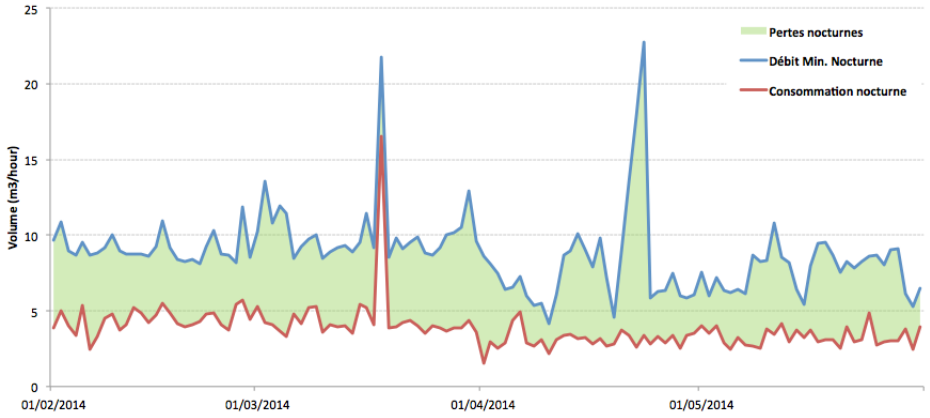


FIGURE 4.2 – Estimation des pertes nocturnes sur le secteur de la Cote 50.

On constate deux fortes variations du débit minimum nocturne, laissant supposer l'apparition de deux fuites. Cependant, si on dessine en parallèle la courbe des consommations nocturnes estimées, on constate que l'un des deux pics n'est pas dû à l'apparition d'une fuite mais à des consommations nocturnes exceptionnelles. L'estimation des pertes à partir d'un télérelevé (partiel ou complet) des compteurs d'eau permet non seulement d'évaluer les volumes d'eau perdus dans le réseau, mais peut aussi être un outil d'amélioration des techniques existantes de détection de fuites, notamment la méthode du débit minimum nocturne.

4.2 Une première méthode de détection des fuites

En soi l'information sur les volumes perdus sur le réseau est importante mais ne suffit pas pour une gestion en temps réel des réseaux d'eau. En effet, certaines fuites ne sont pas détectables (*cf.* Figure 1.1) auquel cas il est inutile de lancer une campagne de recherche et de réparation de fuites. Par ailleurs, les pertes estimées prennent en compte différentes sources d'erreur (listées ci-avant) et sont influencées par la pression de service. Ces différents facteurs tendent à bruiter le signal des pertes comme le montrent la Figure 4.3.

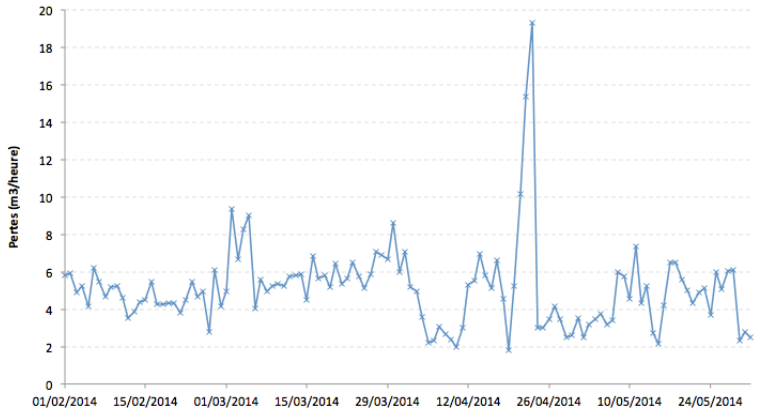


FIGURE 4.3 – Estimation bruitée des pertes nocturnes sur le secteur de la Cote 50.

Il convient alors de créer un outil capable de lisser la courbe des pertes estimées et de pouvoir à partir de cette nouvelle information détecter des fuites, de préférence détectables par l'opérateur.

4.2.1 Cartes de contrôle sur les pertes

Suivant les besoins exprimés ci-avant, il est apparu naturel de construire une carte de contrôle sur les pertes. La carte de contrôle est un outil graphique permettant d'analyser si un processus (dans notre cas les pertes) est sous contrôle (il n'y a pas d'anomalies avérées) et de déclencher une alerte en cas d'évènements atypiques. Le graphe construit correspond au tracé d'une statistique calculée à partir du procédé initial à laquelle viennent s'ajouter des limites de contrôle : une anomalie est détectée à un instant t si cette statistique dépasse les limites de contrôle au même temps t. Il existe dans la littérature différents types de contrôle (carte de Shewhart, carte CUSUM, etc.) ; nous nous intéresserons ici à une carte de contrôle en particulier : la carte EWMA [Roberts, 1959].

La carte EWMA (Exponentially Weighted Moving Average) calcule une statistique (notée $R(t)$) à partir de la valeur actuelle du processus pondérée par les valeurs historiques. Soit $L(t)$ la valeur au temps t du processus d'intérêt, la statistique $R(t)$ se calcule de la façon suivante :

$$R(t) = \lambda L(t) + (1-\lambda)R(t-1) \tag{4.1}$$

où $0 < \lambda \leq 1$ est un paramètre permettant de moduler la mémoire du processus $R(t)$. La carte EWMA pour $\lambda = 1$ correspond à une carte de Shewhart ; pour $\lambda \to 0$, la carte correspond à une carte CUSUM. On définit par ailleurs une valeur de départ pour le processus,

$R(0) = \mu_0$, valeur pour laquelle on dira que le processus est sous contrôle statistique, i.e. il ne manifeste aucune anomalie. Nous construisons un intervalle de fluctuation autour de ce seuil, intervalle dans lequel le processus peut évoluer sans que l'on déclenche une alarme. Cette intervalle permet ainsi de prendre en compte l'imprécision du processus initial $L(t)$. Ces limites de contrôle (LCL et UCL) sont construites à partir de la valeur seuil μ_0 ainsi que de l'écart-type du processus $R(t)$ (noté σ_R), lui même défini à partir de l'écart-type du processus initial (noté σ_L). [Roberts, 1959] définit une valeur limite de ces bornes, pour des mesures indépendantes et identiquement distribuées dont la variance est connue et constante, qui vaut :

$$[LCL\,;UCL] = \mu_0 \pm k\,\sigma_R = \mu_0 \pm k\,\sigma_L\sqrt{\frac{\lambda}{2-\lambda}} \tag{4.2}$$

où k est un coefficient d'élargissement de l'intervalle de contrôle. Une fois les différentes variables définies, le principe de la carte de contrôle est assez simple :

$$\begin{cases} R(t) \in [LCL;UCL], & \text{le processus est sous contrôle statistique,} \\ R(t) \notin [LCL;UCL], & \text{une alerte est déclenchée.} \end{cases}$$

Les paramètres λ et k peuvent être définis selon l'ARL (Average Run Length). L'ARL(δ) est le temps moyen nécessaire à la carte pour détecter une anomalie d'ordre $\delta\sigma_R$. L'ARL(0) est ainsi le temps moyen que mettra la carte à déclencher une alarme sans qu'il y ait d'anomalie avérée (faux positif). Il est évident que l'on cherche à maximiser l'ARL pour $\delta = 0$ et au contraire le minimiser pour $\delta > 0$. [Lucas and Saccucci, 1990] définissent un ensemble de couples (λ, k) optimaux en fonction d'un ARL(0) et d'un δ désirés.

Pour appliquer les cartes de contrôle à la détection de fuites sur le réseau, il faut redéfinir dans notre contexte l'ensemble des variables présentées précédemment. Le processus $L(t)$ représente les pertes journalières calculées par différence entre les VLAR journaliers et l'estimation du total de la consommation journalière. Concernant la précision de ce processus, nous faisons une (forte) hypothèse sur le fait que la précision de $L(t)$ ne dépend que de l'écart-type de l'estimateur des consommations. En effet, les erreurs de mesure (des appareils de mesure d'entrée et des compteurs de vente d'eau) peuvent être considérées comme négligeables face à l'imprécision des consommations estimées, nous prenons ainsi $\sigma_L = \sqrt{\mathrm{Var}(\hat{T}_Y(t))}$.

Le seuil μ_0 représente la valeur du processus hors anomalie. Nous pourrions fixer la valeur de ce seuil à 0, sauf qu'il n'existe aucun réseau pour lequel les pertes journalières sont nulles. En effet, malgré une bonne gestion du réseau, il reste des fuites non visibles et non détectables sur lesquelles l'opérateur a peu de moyen d'action (*cf.* Figure 4.4).

Ce volume de fuites, nommé UBL (Underground Background Leakage), peut être approché d'après [Melato et al., 2009]) en utilisant l'équation suivante

$$\text{UBL (L/h)} = (20 \times Lm + 1.25 \times Ns) \times \left(\frac{AZNP}{50}\right)^{1.5}. \tag{4.3}$$

où Lm est le linéaire de canalisation en kilomètre, Ns est le nombre de branchements et $AZNP$ est la pression moyenne nocturne, en mètre de colonne d'eau. Ces données permettent d'estimer une valeur de μ égale à 132 m^3/jour pour Canéjan et de 4.5 m^3/h pour la Cote 50.

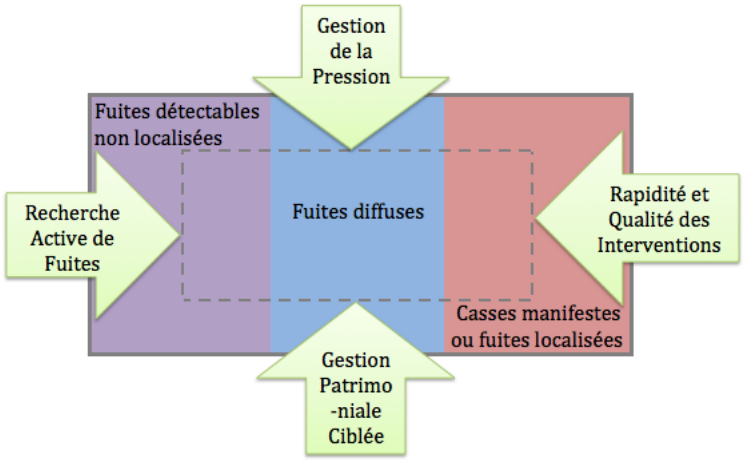


FIGURE 4.4 – Les trois états de fuites et l'impact des pratiques de gestion des réseaux d'eau.

Enfin pour déterminer les valeurs de λ et k, il faut au préalable fixer l'ARL(0) ainsi que le δ désirés. Sachant qu'en moyenne le temps de ré-apparition d'une fuite est entre 2 mois et 2 mois 1/2 (sur l'agglomération bordelaise aux dires des exploitants), un ARL(0)=100 est largement suffisant. Cela signifie qu'il faudra en moyenne 100 jours au modèle à détecter un faux positif alors qu'en 100 jours une fuite sera de toute façon apparue sur le réseau. Le δ que nous nous fixons est égal à 1, l'écart-type de l'estimateur des consommations étant supposé être égal au débit d'une fuite invisible, nous souhaitons pouvoir détecter des écarts de l'ordre de σ_R. [Lucas and Saccucci, 1990] définissent ainsi $\lambda \in [0.16; 0.19]$ et $k \in [2.298; 2.346]$.

Ces données permettent de construire la carte de contrôle présentée à la Figure 4.5 (pour $\lambda = 0.175$ et $k = 2.3$). La carte de contrôle EWMA a été réalisée à l'aide du package R *qcc* [Scrucca, 2004]. Dans notre cas d'étude, nous ne nous intéresserons qu'aux alertes déclenchées par le cas $R(t) > UCL$.

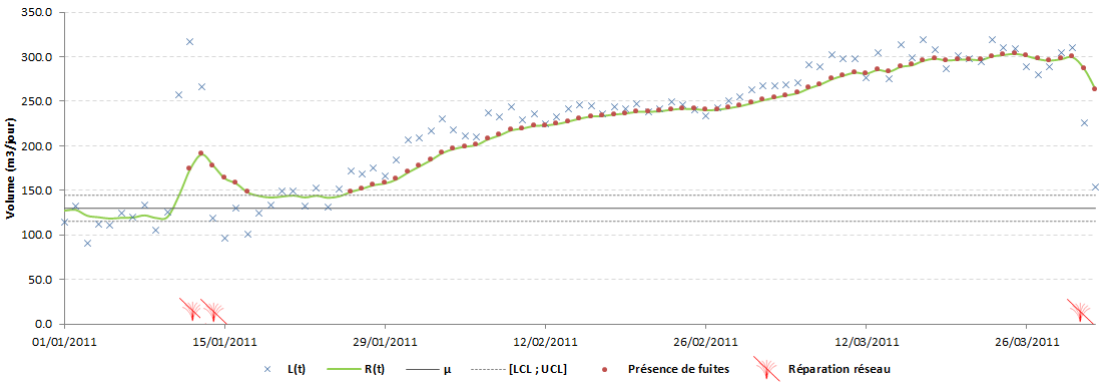


FIGURE 4.5 – Carte de contrôle EWMA pour la détection des fuites sur le secteur de Canéjan

Sur la Figure 4.5, en plus de la carte de contrôle, nous avons indiqué les interventions enregistrées sur le réseau. Deux évènements ont eu lieu :

– deux réparations de casses sur un branchement les 13 et 14 janvier 2011,
– une réparation d'une fuite invisible sur une canalisation le 31 mars 2011.

On remarque, pour le premier évènement, la rapidité de détection de la fuite par le modèle mais aussi par l'opérateur (la fuite est visible en surface). Pour le second évènement, on peut supposer l'apparition de l'anomalie aux alentours du 23 janvier alors que la réparation n'a lieu que deux mois plus tard, le 31 mars. La carte de contrôle semble être plus rapide puisque qu'elle déclenche une alerte au 26 janvier soit seulement trois jours après l'apparition supposée de l'évènement. Il faut noter que la carte de contrôle utilise des données historiques ; le tracé de la courbe n'aurait pas été identique si la carte de contrôle avait été utilisée en temps réel, car alors une alerte aurait été déclenchée, la fuite aurait été plus rapidement réparée et la trajectoire de la courbe vraisemblablement changée.

Le même procédé est mis en œuvre à partir des données de volumes de pertes nocturnes sur la Cote 50 présentées à la Figure 4.2. Les paramètres λ et k étant définis comme pour le cas de Canéjan, la carte de contrôle renvoie les résultats présentés à la Figure 4.6.

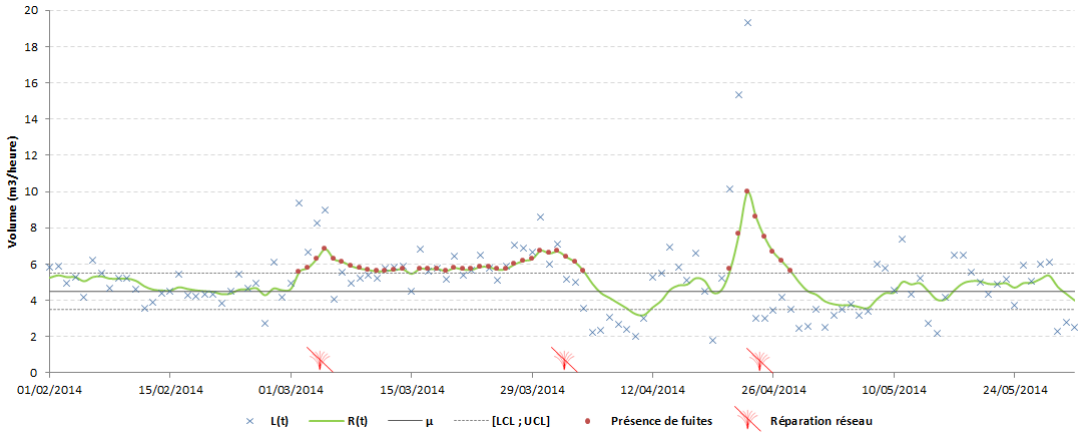


FIGURE 4.6 – Carte de contrôle EWMA pour la détection des fuites sur le secteur de la Cote 50

Trois interventions ont été enregistrées sur le réseau durant cette période :

– deux réparations de branchement suite à une recherche active de fuite les 06/03/2014 et 02/04/2014,
– la réparation d'une casse sur canalisation le 24/04/2014.

On remarque au travers de la carte de contrôle que la première réparation sur branchement n'a pas permis de ramener le volume de pertes à son état initial puisqu'une seconde fuite, réparée pratiquement un mois plus tard, semble expliquer la hausse des pertes.

Ces deux exemples montrent l'importance du suivi des pertes au cours du temps, et non pas uniquement du débit minimum nocturne. La gestion du réseau grâce au télérelevé (même partiel) permet une plus grande réactivité sur la réparation de nouvelles fuites

survenues. Le lissage des courbes de pertes grâce à des méthodes comme les cartes de contrôle réduit par ailleurs les bruits et informations erronées (dus en partie à l'imprécision sur les données de débit entrant et de consommation).

4.2.2 Les limites des cartes de contrôle

L'utilisation d'une carte de contrôle EWMA est soumise à différentes hypothèses concernant le processus de départ. Tout d'abord, les cartes de contrôle supposent que le processus suit une loi normale, ce que nous ne garantissons pas dans notre étude. Cependant, il est à noter que les cartes EWMA ne sont pas sensibles à cette hypothèse du processus initial ([Montgomery, 2009]). Il existe cependant d'autres hypothèses auxquelles la carte EWMA est sensible, notamment le fait que les données du processus initial ne doivent pas être auto-corrélées et que le processus doit être homoscédastique.

Dans la mesure où nous traitons des pertes sur le réseau, il est évident que la perte en eau à un temps t dépend des pertes antérieures (hors intervention de l'exploitant). Ainsi, non seulement nous ne pouvons garantir l'indépendance des données, mais nous sommes même certains du contraire. Notons que l'auto-corrélation n'a pas d'effet sur la construction des cartes pour une valeur d'auto-corrélation du premier ordre inférieure à 0.8 [Woodall and Faltin, 1993]. L'utilisation de données auto-corrélées entraîne des problèmes de différentes natures :

- On peut tout d'abord spécifier que l'application directe d'une carte de contrôle à des données auto-corrélées entraîne un biais dans l'estimation des paramètres ([Reynolds and Chao-Wen, 1997]) tels que λ dans le cas EWMA.
- De manière générale, le vrai ARL (nombre d'observations avant de détecter un changement dans le processus) est sous-estimé si on considère à tort l'indépendance entre les variables ([Schmid and Schone, 1997]).
- Si l'auto-corrélation est positive, les observations ont tendance à prendre les mêmes valeurs, ce qui implique une dérive du processus au cours du temps [Woodall and Faltin, 1993]. Cela entraîne alors une augmentation du taux de fausses alarmes.
- Si l'auto-corrélation est négative, les observations successives tendent à prendre des valeurs dissemblables. On se retrouve avec un schéma en "dent de scie". Le risque lié à l'utilisation d'une carte de contrôle traditionnelle est une lente détection des changements du processus.

La seconde limite de l'utilisation des cartes de contrôle à notre cas d'étude est l'hétéroscédascticité du processus. Il est difficilement acceptable de prétendre que la variance (ou l'écart-type) du processus est constante dans le temps comme le montre la Figure 3.14. Les cartes de contrôle classiques (Shewart, CUSUM ou EWMA), qui ne prennent pas en compte l'hétéroscédasticité d'un processus, ne peuvent pas fournir d'informations fiables sur l'état de celui-ci ; les limites de contrôle conventionnelles sont invalides.

Il existe cependant des solutions palliatives à ces deux problèmes. L'auto-corrélation des données peut être facilement contournée par l'utilisation de séries temporelles. Le processus peut-être modélisé par un modèle ARMA (*AutoRegressive Moving Average*, [Alwan and Roberts, 1988]) et la carte de contrôle est appliquée sur les résidus ([Croux et al., 2011]). Concernant l'hétéroscédasticité du processus, [Fang and Zhang, 1999] pré-

conise la modélisation du processus suivant un modèle GARCH (*General AutoRegressive Conditional Heteroscedastic*). Pour prendre en compte conjointement les deux problèmes, l'utilisation d'un modèle ARMA-ARCH est recommandée : le processus L_t est modélisé selon un processus ARMA et un modèle ARCH est appliqué sur les résidus. La carte de contrôle appliquée sur ces derniers a des limites variant dans le temps.

4.2.3 Vers une meilleure modélisation des fuites et des ouvrages d'eau potable

Au vu de la complexité des modèles à développer d'une part et compte tenu de l'objectif final, nous n'avons pas développé l'analyse des cartes de contrôle pour les pertes en eau potable. L'objectif de ces cartes était de fournir un outil, facilement lisible pour des gestionnaires d'eau potable, permettant de suivre l'évolution des pertes sur le réseau. Cependant, construire une carte de contrôle sur les résidus d'un modèle ARMA risque de complexifier la lecture de la carte, dans la mesure où la grandeur mesurée et suivie n'a aucune signification opérationnelle. Par ailleurs, nous avons besoin de développer un modèle qui permet de décomposer les fuites en une partie détectable et une partie indétectable afin d'optimiser le pilotage du réseau. C'est pourquoi nous allons préférer un modèle multi-états qui permettra de décomposer le débit de fuite en différents états, mettant alors en avant les différents niveaux de dégradation des fuites et du patrimoine.

4.3 La dégradation des fuites et du patrimoine

Les pertes sur les réseaux d'eau sont en grande partie expliquées par la dégradation du réseau. La dégradation des ouvrages d'eau potable (en particulier les canalisations et les branchements) est due d'une part au vieillissement de ces ouvrages mais aussi à des facteurs endogènes (longueur, diamètre, matériau de l'ouvrage) ou exogènes (pression de l'eau, nature du sol, etc.). Parallèlement à ce phénomène, les fuites sur le réseau s'amplifient au cours du temps, le débit de la fuite (volume d'eau perdue par unité de temps) tendant à s'accroître au cours du temps.

4.3.1 Les états de dégradation

D'après le concept BABE (Burst And Background Estimates, [Lambert, 1994]), il est possible de distinguer trois états de dégradation pour une fuite (*cf.* Figure 1.1), qui dépendent du débit de celle-ci et indirectement de la capacité de l'opérateur à la détecter. Ces trois états pour une fuite sont :

- les fuites diffuses (invisibles et indétectables),
- les fuites détectables (non visibles en surface mais détectable par recherche de fuite),
- les casses manifestes (visible en surface).

Des travaux ont déjà été menés sur la modélisation de l'évolution des fuites selon ces trois états ([Chesneau, 2006]). Ces travaux ont permis, à partir du débit de fuite global, d'associer à chaque état, précédemment mentionné, le volume perdu. En réalité, ce

modèle traduit le débit de fuite global comme une somme pondérée des débits unitaires par état et du nombre de fuites par état. Cependant, par souci de convergence des modèles (convergence pour l'estimation des paramètres), ces travaux sont basés sur des hypothèses qui ne sont pas forcément en concordance avec la réalité.
En tout premier lieu, le modèle ne fait pas de distinction entre les fuites sur canalisation et les fuites sur branchement, alors que les débits associés diffèrent. [Lambert et al., 1999] définit, pour les réseaux d'eau potable, l'UARL (Unavoidable Annual Real Losses) : il s'agit du volume d'eau qui reste perdu malgré une gestion "parfaite" du réseau. La Table 4.1 permet de calculer l'UARL d'un réseau et reprend les valeurs moyennes de débit.

TABLE 4.1 – Valeurs des paramètres utilisées pour le calcul de l'UARL

Ouvrage	Fuites diffuses	Fuites détectables	Casses manifestes
Canalisation	20 L/km/h*	0.006 fuites/km/an à 6 m^3/h*	0.124 casses/km/an à 12 m^3/h*
Branchement jusqu'au bord de chaussée	1.25 L/brt/h*	0.75/1000 brt/an 1.6 m^3/h*	2.25/1000 brt/an 1.6m^3/h*
Branchement après bord de chaussée	0.50 L/brt/h*	0.50/1000 brt/an 1.6 m^3/h*	1.5/1000 brt/an 1.6m^3/h*

* tous les débits sont calculés pour une pression moyenne de 5 bars

On constate ainsi que les débits de fuite liés aux canalisations ou branchements ne sont pas similaires ; ne pas distinguer les deux types d'ouvrages est source d'erreurs potentielles. Par ailleurs, le nombre de casses manifestes est considéré comme négligeable. En effet, dans la mesure où le pas de temps utilisé dans ces travaux est la semaine, une fuite qui surface à un instant t et immédiatement réparée à ce même instant t. Au vu des débits présentés à la Table 4.1, les volumes liés aux casses manifestes ne sont pas négligeables, en particulier si le pas de temps est réduit en deçà de la semaine. Enfin, une dernière critique qui peut être faite sur la modélisation de la dégradation des fuites est la spécification de certains paramètres. Afin de faciliter l'estimation des paramètres du modèle, des degrés de liberté ont été supprimés. Le débit lié aux fuites diffuses et le paramètre associé au taux d'apparition de ces fuites sont fixés au préalable. Enfin on peut noter, en dernier lieu, l'absence de covariables permettant d'expliquer la dégradation des réseaux et l'apparition des fuites dans chacun des états.

Ces modèles, malgré la riche information qu'ils peuvent fournir, sont difficiles à implémenter. En effet, le nombre de fuites dans le second état est partiellement connu (uniquement lors de campagne de recherche de fuites, sur un sous-ensemble du réseau) et le nombre de fuites dans le premier état est totalement inconnu. Nous pouvons cependant mettre en relation les états de dégradation des fuites avec les ouvrages d'eau potable (canalisation et branchement). En effet, ces fuites ont lieu sur ces ouvrages, c'est donc eux qui subissent en réalité une dégradation. On considère dans ce cas, en plus des trois états de fuites, un état de bonne condition de l'ouvrage, où il ne subit aucune fuite. Mais

comme nous l'avons souligné précédemment, il est difficile, voire impossible, d'identifier les fuites diffuses ; nous allons donc fusionner cet état avec l'état initial. Dans la réalité, il n'est pas aberrant de ne pas dissocier ces états, dans la mesure où les fuites diffuses ne pouvant pas être détectées et réparées, l'opérateur n'a aucun moyen d'intervention si ce n'est le renouvellement de l'ouvrage.
Nous considérons ainsi trois états de dégradation des ouvrages d'eau potable :

État 1 (E_1) : l'ouvrage ne subit aucune fuite ou des fuites diffuses,
État 2 (E_2) : l'ouvrage subit des fuites détectables non manifestes en surface,
État 3 (E_3) : l'ouvrage subit des casses manifestes.

On suppose que tous les ouvrages en service sont obligatoirement dans un de ces états. On peut aussi supposer que, dans un temps infinitésimal, un individu peut naturellement soit rester dans un état soit passer dans l'état adjacent supérieur. Au moment de leur pose (naissance de l'ouvrage), les ouvrages sont tous dans l'état E_1. Ainsi d'après les deux dernières hypothèses, on suppose qu'un ouvrage passe successivement par les trois états de dégradation. Notre modèle multi-états peut se représenter schématiquement, voir la Figure 4.7.

FIGURE 4.7 – Modélisation multi-états de la dégradation des ouvrages d'eau potable

4.3.2 Observation de l'état des ouvrages

Le calage du modèle que nous souhaitons mettre en place nécessite de connaître (totalement ou partiellement) les états des ouvrages. L'attribution des états est réalisée à partir des données d'intervention réseau qui permettent de savoir ponctuellement dans quel état se trouve un ouvrage. Les ouvrages dans l'état E_3 sont exhaustivement identifiés à partir des données d'intervention de l'exploitant.
Concernant les états en E_1 et E_2, nous n'avons qu'une vision partielle des ouvrages dans ces états. Lorsqu'une campagne de recherche de fuites est réalisée sur un secteur, il est possible d'identifier clairement, durant la durée de cette campagne, l'ensemble des ouvrages se situant dans l'état E_2 et nous pouvons affirmer que l'ensemble des ouvrages pour lesquels aucune fuite n'a été détectée se situe dans l'état E_1. De fait les observations des ouvrages dans les états E_1 et E_2 ne sont qu'un échantillon spatio-temporel de l'intégralité des ouvrages : on parle alors de données de panel.

Les données d'intervention intègrent donc :
- l'intégralité des ouvrages ayant subi une casse (hors casses provoquées par des entreprises tierces) sur le réseau pendant la période d'observation (E_3),
- un échantillon d'ouvrage subissant des fuites invisibles, fuites détectées suite à une recherche de fuite (E_2)

– un échantillon d'ouvrage, correspondant à un secteur inspecté (recherche de fuite) pendant une période donnée, pour lequel aucune fuite n'a été détectée (E_1),

Les observations des ouvrages dans les états E_1 et E_2 sont donc conditionnées par la recherche de fuite. La carte 4.8 illustre l'inégale répartition des campagnes de recherche de fuites sur le territoire bordelais. Cette inégalité s'explique par le fait que les secteurs ne présentent pas tous une même proportion de fuites invisibles localisables. Le secteur de Paulin Becquet (qui correspond au centre ville de la métropole bordelaise) est celui pour lequel l'effort de recherche de fuites est le plus important.

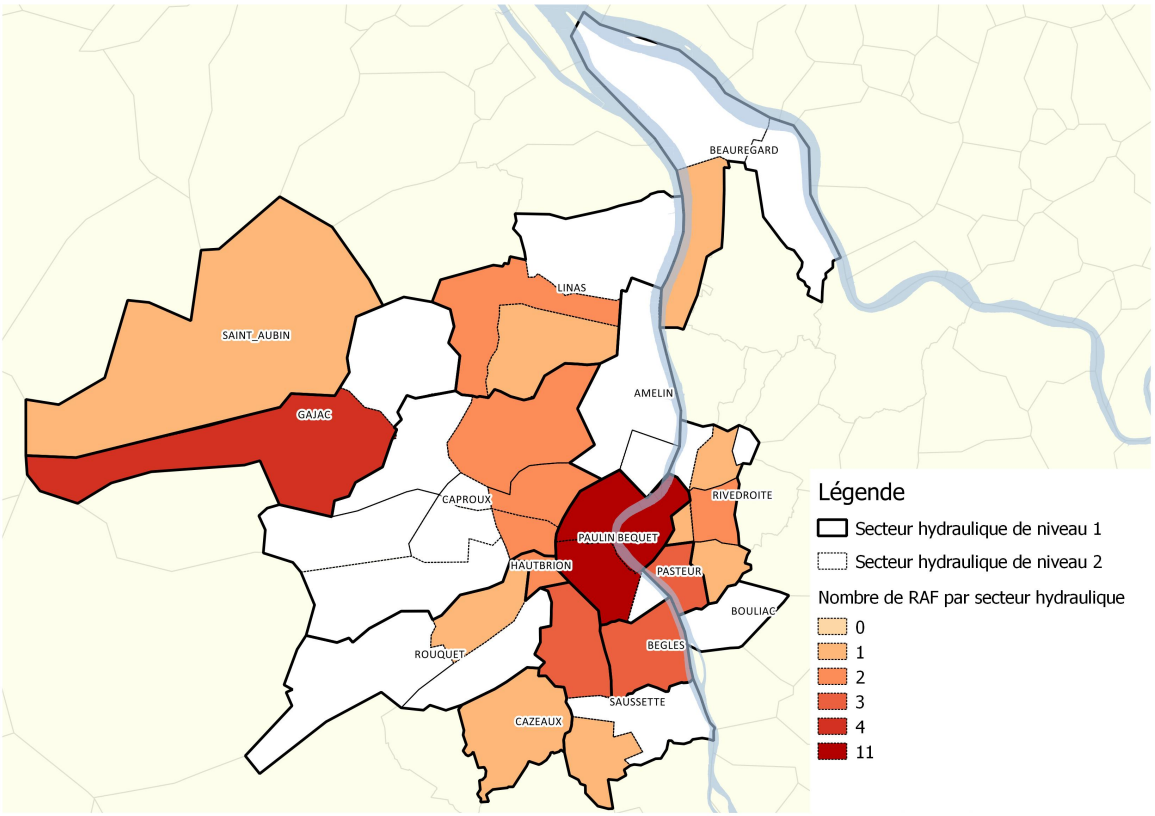


FIGURE 4.8 – Carte des campagnes de recherche de fuite sur la CUB entre 2012 et 2013 (source : Lyonnaise des Eaux)

4.4 Le modèle de pertes en eau

L'objectif du modèle que nous allons construire est de pouvoir dissocier dans le volume de pertes en eau global quelle part est due aux fuites diffuses, aux fuites invisibles et aux casses manifestes. Il est donc possible de distinguer deux phénomènes à modéliser : d'abord la dégradation des ouvrages d'eau potable selon les trois états présentés à la Figure 4.7, puis le débit de perte en eau que nous pouvons associer à chacun de ces états. Le modèle de fuite se scinde alors en deux sous-modèles :

– un premier modèle qui permet d'estimer pour chaque ouvrage un vecteur **P** de probabilités d'être dans chacun des états E_1, E_2 ou E_3. Ce modèle sera calé sur les données de la CUB, à partir de l'intégralité des données de réparation (suite à une casse) ainsi qu'à partir d'un échantillon spatio-temporel d'ouvrages ayant été inspectés.
– une fois ce modèle calé, nous nous focalisons sur une zone où nous disposons des données de pertes journalières. Nous calculerons les probabilités d'appartenance à chaque état et à partir du débit de fuite global, nous calerons une fonction $Q_j(t)$ qui permettra d'estimer le débit de fuite dans chaque état j.

Ces deux modèles seront calés sur les données de la CUB, du fait de la précision des données dont nous disposons, aussi bien sur les données patrimoniales que sur les données d'interventions sur le réseau mais aussi sur celles issues de l'instrumentation de ce réseau (volume livré au réseau, mesure en continu de la pression). Le tableau 4.2 résume, pour l'ensemble des ouvrages considérés de la CUB, la répartition par matériau. On constate une inégale répartition des matériaux sur le réseau d'eau potable de la CUB, aussi bien pour les branchements que pour les canalisations.

TABLE 4.2 – Répartition du patrimoine d'eau potable (canalisations et branchements) par nature de matériau sur la CUB (juillet 2014)

Matériau	**% Linéaire canalisation**	**% Nombre branchements**
Inconnu		2%
Acier	<1%	
Amiante ciment	<1%	
Divers	<1%	2%
Fonte ductile	28%	<1 %
Fonte grise	55%	<1%
Polyéthylène bleu	2%	76%
Polyéthylène noir	<1%	13%
Plomb	<1%	3%
PVC	14%	4%

De façon générale, il est constaté que les matériaux ne vieillissent pas de la même façon et réagissent différemment aux facteurs externes (pression, agressivité des sols, conditions climatiques, etc.). Il est alors naturel de construire un modèle par nature de matériau pour les principaux matériaux du réseau.

Concernant les canalisations, nous construirons trois modèles : fonte grise, fonte ductile et PVC. Ces matériaux représentent plus de 97% du linéaire du réseau d'eau potable de la CUB (avec 83% de fonte). Les matériaux très peu représentés sur le réseau tels que l'amiante-ciment (0.2% du linéaire), l'acier (0.6% du linéaire) et peu représentés comme le polyéthylène (seulement 2% du linéaire) sont respectivement assimilés aux conduites en fonte grise, fonte ductile et PVC. Trois familles de matériaux ont ainsi été constituées :
– la fonte grise et l'amiante ciment (FG/AM),

– la fonte ductile et l'acier (FD/AC),
– le PVC, le polyéthylène (PE) bleu et le PE noir (PVC/PE).

Les autres matériaux ne sont pas modélisés.

Pour les branchements, nous remarquons clairement que le polyéthylène est le matériau majoritairement présent. La réglementation (directive européenne 98/83/CE) poussant à limiter la présence du plomb dans les réseaux d'eau potable, les branchements en plomb disparaitront des réseaux publics d'eau potable, en particulier sur la CUB à la fin de l'année 2014. Concernant les branchements en PVC, compte tenu de leur faible effectif, il est difficile de modéliser ce matériau et il n'est pas possible de les rattacher à des branchements en polyéthylène, contrairement aux canalisations, du fait d'un vieillissement évoluant différemment. Les autres matériaux (notamment le cuivre ou l'acier par exemple) sont eux aussi trop faiblement présents pour être exploitables dans les modèles. La modélisation des branchements se portera donc uniquement sur les seuls branchements en polyéthylène (PE noir et PE bleu).

Après avoir filtré notre base de données sur les matériaux qui seront modélisés, et regroupé les matériaux entre eux, nous obtenons la répartition du patrimoine et des inspections suivante (*cf.* Tableau 4.3).

TABLE 4.3 – Répartition du patrimoine et des inspections sur le réseau d'eau potable de la CUB

Ouvrage	Materiau	Effectif*	Nombre d'inpection*
Branchement	PE bleu	100 627	489 182
	PE noir	17 582	52 529
Canalisation	FD/AC	857	2 391
	FG/AM	1 656	6 387
	PVC/PE	485	1 165

* en nombre pour les branchements, en linéaire (km) pour les canalisations.

La Figure 4.9 présente la répartition par états des ouvrages en fonction du matériau ; cette repartition est exprimée en % du nombre d'ouvrages inspectés dans un état pour les branchement et en % du linéaire inspecté pour les canalisations.

4.5 Le modèle de dégradation des ouvrages

Il existe déjà des outils opérationnels modélisant la dégradation des ouvrages : certains traitent de la défaillance ponctuelle des ouvrages (comme le LEYP que nous présenterons par la suite) et d'autres s'intéressent à l'évolution des ouvrages à travers différents états de dégradation. C'est le cas du modèle GompitZ [Le Gat, 2008], qui est appliqué au réseau d'assainissement. Ce modèle opérationnel a toutefois certaines limites notamment par le fait qu'il n'intègre pas la dépendance entre les différents états. C'est pourquoi nous décidons de développer un modèle multi-états semi-markovien [V.Barbu and N.Limnios, 2008]. Ce modèle permet de modéliser le passage d'un ouvrage à travers différents états de dégradation qui mèneront à l'apparition d'une casse manifeste. Une des bases communes de ces deux modèles est l'utilisation d'un facteur de type Cox sur l'intensité du

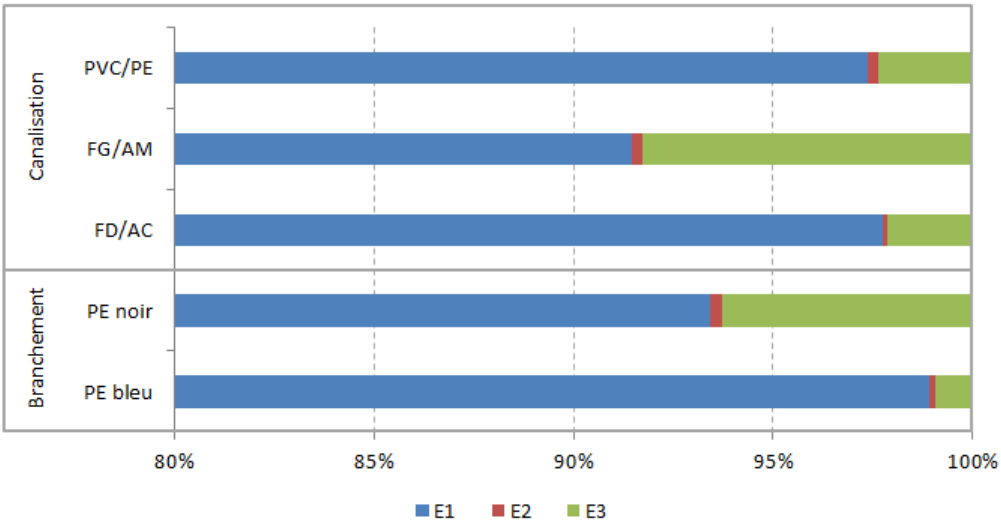


FIGURE 4.9 – Répartition des inspections selon les 3 états par nature de matériau.

processus stochastique, permettant d'intégrer l'effet de covariables propres à l'ouvrage, à son environnement ou à son fonctionnement.

Le modèle utilisé sera calé sur les données patrimoniales de la Communauté Urbaine de Bordeaux ; les données d'intervention utilisées couvrent la période 2000-2013.

4.5.1 Covariables a priori explicatives de la dégradation des ouvrages

La dégradation des ouvrages d'eau potable est la conséquence des différents stress que subit l'ouvrage et qui peut s'expliquer par des facteurs intrinsèques à celui-ci mais aussi par des facteurs externes. Nous présentons ici une liste (non exhaustive) des facteurs explicatifs du risque d'apparition et de dégradation des fuites sur les ouvrages que nous utiliserons dans notre modèle multi-états. Il est important de noter qu'étant donné les divergences dans la dégradation des ouvrages mais aussi dans la disponibilité des données, le modèle sera stratifié par nature d'ouvrage (canalisation ou branchement).

Matériau
Les différents modèles statistiques modélisant les défaillances sur les conduites d'eau potable (voir par exemple [Eisenbeis, 1994] ou [Le Gat, 2009]) sont stratifiés par nature de matériau. Il est naturel de supposer a priori un effet des facteurs explicatifs différent d'un matériau à un autre, chaque type de matériau réagissant de façon propre aux agressions physico-chimiques de son environnement. Cette hypothèse sera validée si les paramètres ont des plages de valeurs spécifiques à chaque matériau.

Longueur
La longueur d'un composant a un effet aggravant le risque de défaillance. Certains auteurs considèrent que le taux de défaillance est proportionnel à une puissance de la longueur, variable selon le matériau. [Andreou, 1986, p.166] suggère même que le risque de défaillance est approximativement proportionnel à la racine carré de la longueur.

Diamètre
Différentes études (voir par exemple [Eisenbeis, 1994]) ont montré qu'un grand nombre de défaillances apparaissaient sur les ouvrages de petit diamètre. Les conduites ayant un diamètre inférieur ou égal à 200 mm sont particulièrement sujettes à la casse. Cette sensibilité peut s'expliquer par le fait que la résistance de la conduite est réduite du fait d'une paroi plus fine pour les petits diamètres. Par ailleurs, une autre raison peut venir du faible débit de l'eau qui y circule résultant alors d'une stagnation des matières en suspension créant ainsi un environnement favorable pour le développement des bactéries. Enfin, les conduites de gros diamètres étant plus lourdes que celles de petit diamètre, cela implique une plus grande précaution lors de l'installation des conduites.

Période de pose
La période de pose est un facteur regroupant différents phénomènes, principalement la technologie de matériau utilisée. Cette variable est en définitive propre à chaque exploitation et peut difficilement se généraliser à l'ensemble des réseaux. La principale source de renseignement est donc l'expert/opérateur qui a une connaissance de l'histoire du réseau et des ouvrages installés. Sur le réseau de la Communauté Urbaine de Bordeaux, nous distinguons différentes périodes de pose en fonction des matériaux :

- Fonte Grise/Amiante Ciment (canalisation) : nous distinguons les conduites posées avant 1934 (meilleures conditions de pose) de celles posées après 1934. Ceci est principalement dû à la méthode de fabrication des conduites : avant les années 1930, le procédé de fabrication était la coulée (horizontale ou verticale) alors qu'après les années 1930, la centrifugation a été utilisée, les conduites étant alors moins épaisses qu'auparavant.
- Polyéthylène Noir (branchement) : la première génération de PE Noirs (posée avant 1977) est plus fragile que les générations suivantes.
- Polyéthylène Bleu (branchement) : la première génération de PE Bleus (posée entre 1986 et 1990) est plus fragile que les générations suivantes.

Nombre de casses historiques
La répétition d'incidents sur une conduite accélère son vieillissement [Le Gat and Eisenbeis, 2000]. Goulter and Kanzemi [1988] suggèrent même que les casses répétées sont conséquentes aux dommages occasionnés pendant la réparation de la conduite, tels que les "coups de bélier" (augmentation brutale de la pression) au moment de la réhabilitation de la conduite ou encore les mouvements de sols causés par l'excavation, le rebouchage et le trafic de véhicules lourds.

Pression
La pression a un effet avéré sur le risque de défaillance et la dégradation des ouvrages ([Van Zyl and Clayton, 2007]). L'effet qui relie la pression et la fuite diffère en fonction de la nature du matériau. Lambert [2001] établit même une relation entre la variation de pression (P_1/P_0) et la variation du débit de fuite (L_1/L_0) où

$$\frac{L_1}{L_0} = \left(\frac{P_1}{P_0}\right)^{N_1}.$$

L'exposant N_1 peut varier entre 0.5 et 2.5 en fonction du type de fuite ou du type de maté-

riau (rigide, type fonte, ou plastique, type polyéthylène). Nous utiliserons dans ce modèle la pression statique (pas de variation au cours du temps) qui est obtenue pour chaque ouvrage à partir d'un modèle hydraulique. Nous disposons ainsi d'une pression minimale $P_{\min}$, d'une pression maximale $P_{\max}$ et d'une variation de pression Var_P (=$P_{\max}$-$P_{\min}$) pour chaque ouvrage.

Sol argileux
Morris [1967] et Clark [1971] soulignent le fait que le retrait-gonflement des argiles est un facteur contribuant à la dégradation des ouvrages. Nous utiliserons alors une variable indicatrice de la pose d'un ouvrage dans un sol argileux.

Conditions de pose
Les facteurs liés aux conditions de pose des ouvrages sont rarement développés dans la littérature du fait qu'ils dépendent uniquement de l'exploitation du réseau. De fait, les variables liées aux conditions de pose exploitées dans nos modèles sont issues du retour d'expérience des exploitants sur le réseau de la CUB. Nous identifions trois facteurs en particulier :

- le lotissement : ce facteur concerne essentiellement les branchements. Les branchements posés dans un lotissement sont installés par le lotisseur au moment de la construction et non pas par l'opérateur. On suppose alors que la pose des branchements dans un lotissement est de moins bonne qualité que pour les branchements posés par l'opérateur sur le domaine public. Le calage du modèle et le test de significativité des paramètres permettra de valider ou d'invalider cette hypothèse,
- le collier de prise en charge : cette variable concerne uniquement des branchements en polyéthylène. Le collier de prise en charge permet de connecter le branchement à une canalisation. Entre 2000 et 2008, une nouvelle génération de collier est apparue. L'exploitant ayant une mauvaise connaissance sur les conditions d'installation de ces colliers, leur pose a entraîné une série de fuites associées aux branchements,
- le fourreau : à partir de 2008, les branchements posés en tranchée ouverte sont installés dans un fourreau assurant la protection et permettant le remplacement rapide du branchement.

Climat
L'effet du climat est complexe et souvent indirect. On suppose très fortement que le froid a un effet sur les conduites en métal. [Kleiner and Rajani, 2002] construisent un *Freezing index* : il s'agit d'un indicateur (calculé en degré-jour) de nombre de jours durant une période où la température est inférieure à un certain seuil τ fixé. Cette variable a été choisie après avoir observé que des pics de casses survenaient périodiquement à la fin de l'hiver quand la température des sols était la plus basse. Il faut aussi noter que de fortes périodes de temps sec ont tendance à favoriser le retrait-gonflement des argiles ([Wols and Van Thienen, 2014]), les conduites en plastique étant plus sensibles à ce genre de phénomène.

Ces différentes covariables vont former en partie le jeu de données utilisé pour caler et valider nos modèles. La construction de ce jeu de données est le résultat du croisement des différentes bases de données à notre disposition (données patrimoniales, données environnementales, données d'intervention, etc.). Le jeu de données utilisé pour le calage

des modèles prend la forme suivante :

TABLE 4.4 – Exemple de données d'entrée pour le modèle de dégradation

Identifiant	Date de pose	Date inspection	État	Classe	Matériau	···
103/104	01/05/1950	26/10/2012	E_3	CANA	FG/AM	···
191582	07/08/1985	14/02/2013	E_1	BRT	PE Bleu	···
192673	23/04/1996	14/02/2013	E_2	BRT	PE Bleu	···

Les données collectées vont nous permettre de caler les paramètres des modèles et nous pourrons ainsi estimer la probabilité pour un ouvrage d'être dans un état de dégradation E_1, E_2 ou E_3.

4.5.2 Le modèle GompitZ

Considérons un réseau d'eau potable composé de N ($N \in \mathbb{N}$) ouvrages (canalisations et branchements). Considérons aussi qu'un composant i ($i = \{1, \cdots, N\}$) est caractérisé par un ensemble des covariables $\mathbf{Z}_i$. Cet individu i a été observé au total m_i fois à des temps t_{ij} ($j = 1, \cdots, m_i$) dans les états X_{ij} ($X_{ij} \in \{1, 2, 3\}$). Nous définissons O_i comme la synthèse des connaissances disponibles sur le composant i :

$$O_i = \{\mathbf{Z}_i, \{(t_{ij}, X_{ij}), j = 1, \cdots, m_i\}\}. \tag{4.4}$$

Remarque : Le vecteur de covariables $\mathbf{Z_i}$ peut être scindé en deux sous-vecteurs, un premier sous-vecteur $\mathbf{Z}_{i0}$ caractérisant les conditions de pose d'un ouvrage et un second sous-vecteur $\mathbf{Z}_{i1}$ caractérisant sa dégradation au cours du temps.

Fonction de survie du modèle GompitZ

Notons $X(t)$ la dégradation d'un ouvrage au temps t. Un ouvrage peut être soit dans un état non ou peu dégradé ($X(t) = 1$), soit moyennement dégradé ($X(t) = 2$) ou soit fortement dégradé ($X(t) = 3$). La survie dans un de ces états de dégradation s'écrit :

$$S_j(t) = \mathbb{P}\{X(t) \leq j\}, \qquad j \in \{1, 2, 3\}.$$

La fonction de survie, exploitée dans notre modèle, est similaire à celle utilisée dans le cadre de la modélisation de la dégradation des collecteurs d'assainissement ([Le Gat, 2008]), elle est inspirée des modèles Gompertz. A noter qu'il ne s'agit pas d'une survie au sens de l'analyse de survie pour une variable de durée, il n'y a pas ici de variable de durée (nous n'imposons pas que $X(t)$ soit un processus multi-états au sens classique). La fonction de survie pour un individu i, est définie de la façon suivante :

$$S_j(t, \mathbf{Z_i}, \theta) = \exp(-\exp(\alpha_j + Z'_{i0}\beta_0 + t\exp(Z'_{i1}\beta_1))) \qquad k \in \{1, 2\} \tag{4.5}$$

et

$$S_3(t, \mathbf{Z_i}, \theta) = 1.$$

avec $\theta \in \mathbb{R}^p$, $\theta' = (\alpha_1, \alpha_2, \beta_0', \beta_1')$. α_j est le paramètre lié à l'état j définissant les probabilités d'état à l'instant $t = 0$, β_0' est le vecteur des paramètres associés aux facteurs de condition de pose et β_1' le vecteur des paramètres associés aux facteurs liés au vieillissement. Le facteur $\exp(Z_{i1}'\beta_1)$ est strictement positif, interdisant ainsi toute amélioration des conditions de vie des ouvrages avec leur vieillissement.
Nous pouvons, à partir des fonctions de survie, définir les probabilités d'appartenance à chaque état $j \in \xi$:

$$\mathbb{P}_\theta\{X_i(t) = j|\mathbf{Z_i}\} = S_j(t, \mathbf{Z_i}, \theta) - S_{j-1}(t, \mathbf{Z_i}, \theta) \tag{4.6}$$

On remarque à l'équation (4.6) qu'au vu de l'écriture des probabilités d'appartenance à un état, une contrainte doit être imposée sur les paramètres d'état α_j ($j \in \{1, 2\}$). En effet, nous imposons la contrainte $\alpha_1 > \alpha_2$ (*cf.* Encadré 1), permettant d'éviter l'intersection des courbes de survie et d'obtenir des probabilités positives dans l'équation (4.6).

Encadré 1 : Contrainte sur les paramètres d'état α_j

La contrainte imposée sur les paramètres d'état α_j permet de s'assurer que les probabilités définies à l'équation (4.6) soient positives. D'après cette équation, nous pouvons imposer une condition sur le ratio des fonctions de survie :

$$\mathbb{P}_\theta\{X_i(t) = j|\mathbf{Z_i}\} = S_j(t, \mathbf{Z_i}, \theta) - S_{j-1}(t, \mathbf{Z_i}, \theta) > 0 \Leftrightarrow \frac{S_j(t, \mathbf{Z_i}, \theta)}{S_{j-1}(t, \mathbf{Z_i}, \theta)} > 1.$$

Suivant l'écriture explicite des fonctions de survie définies à l'équation (4.5), nous avons :

$$\begin{aligned}
\frac{S_j(t, \mathbf{Z_i}, \theta)}{S_{j-1}(t, \mathbf{Z_i}, \theta)} &= \frac{\exp(-\exp(\alpha_j + Z_{i0}'\beta_0 + t\exp(Z_{i1}'\beta_1)))}{\exp(-\exp(\alpha_{j-1} + Z_{i0}'\beta_0 + t\exp(Z_{i1}'\beta_1)))} \\
&= \exp\big(-\exp(\alpha_j + Z_{i0}'\beta_0 + t\exp(Z_{i1}'\beta_1)) \\
&\quad + \exp(\alpha_{j-1} + Z_{i0}'\beta_0 + t\exp(Z_{i1}'\beta_1))\big) \\
&= \exp\big(\exp(Z_{i0}'\beta_0 + t\exp(Z_{i1}'\beta_1)) \times (\exp(\alpha_{j-1}) - \exp(\alpha_j))\big).
\end{aligned}$$

De fait nous avons

$$\begin{aligned}
\mathbb{P}_\theta\{X_i(t) = j|\mathbf{Z_i}\} > 0 &\Leftrightarrow \frac{S_j(t, \mathbf{Z_i}, \theta)}{S_{j-1}(t, \mathbf{Z_i}, \theta)} > 1 \\
&\Leftrightarrow \exp(\alpha_{j-1}) - \exp(\alpha_j) > 0 \\
&\Leftrightarrow \alpha_{j-1} > \alpha_j.
\end{aligned}$$

Nous avons de fait la condition imposée : $\alpha_1 > \alpha_2$.

Vraisemblance du modèle

La contribution à la vraisemblance d'un individu i avec des observations O_i indépendantes (*i.e* des variables aléatoires $X(t_{ij})$ indépendantes), est :

$$\mathscr{L}_i(\theta|O_i) = \prod_{j=1}^{m_i} \mathbb{P}\{X(t_{ij}) = X_{ij}\} = \prod_{j=1}^{m_i} S_{X_{ij}}(t_{ij}, \mathbf{Z_i}, \theta) - S_{X_{ij}-1}(t_{ij}, \mathbf{Z_i}, \theta) \tag{4.7}$$

En considérant que les N vecteurs d'observations O_i sont indépendantes, la vraisemblance pour le N-uplet $\mathbf{O} = \{O_i, i = 1, \cdots, N\}$ est :

$$\mathscr{L}(\theta|\mathbf{O}) = \prod_{i=1}^{N} \mathscr{L}_i(\theta|O_i). \tag{4.8}$$

Estimation des paramètres

L'estimation du vecteur de paramètres $\theta = (\alpha_1, \alpha_2, \beta_0', \beta_1')$ se fait en maximisant la log-vraisemblance :

$$\begin{aligned} l(\theta) = \ln \mathscr{L}(\theta|\mathbf{O}) &= \ln \left(\prod_{i=1}^{N} \prod_{j=1}^{m_i} \mathbb{P}\{X(t_{ij}) = X_{ij}\} \right) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{m_i} \ln(\mathbb{P}\{X(t_{ij}) = X_{ij}\}) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{m_i} \ln(S_{X_{ij}}(t_{ij}, \mathbf{Z_i}, \theta) - S_{X_{ij}-1}(t_{ij}, \mathbf{Z_i}, \theta)) \end{aligned} \tag{4.9}$$

Nous cherchons le vecteur de paramètre $\hat{\theta}$ tel que :

$$\hat{\theta} = \arg\max_{\theta} \{l(\theta)\}. \tag{4.10}$$

L'utilisation de la fonction *constrOptim* de R ([R Development Core Team, 2011]) permet de calculer le jeu de paramètres estimé $\hat{\theta}$ qui maximise numériquement la log-vraisemblance tout en conservant la contrainte imposée ($\alpha_1 > \alpha_2$). Pour cela, il est nécessaire de calculer le gradient de la fonction $l(\theta)$ définie à la fonction (4.9). Le gradient de la log-vraisemblance pour le modèle GompitZ est donné en Annexe G.

Significativité des paramètres

Une conséquence de l'estimation par maximum de vraisemblance (voir par exemple [Rao, 1965]) est que, sous des conditions classiques de régularité, $\hat{\theta}$ est asymptotiquement gaussien d'espérance θ et de matrice de covariance Σ qui peut être estimée par :

$$\hat{\Sigma} = (-\mathbf{H})^{-1} \qquad \text{avec} \qquad \mathbf{H} = \left(\frac{\partial^2 l(\hat{\theta}|\mathbf{O})}{\partial \hat{\theta}_i \partial \hat{\theta}_j} \right)$$

Pour tester la significativité des paramètres, nous utilisons un test du Khi-2 de Wald. Les statistiques présentées ci-après sont distribuées selon une loi du Khi-2 à 1 degré de liberté

$$\forall a \in \mathbb{R}, \forall i \in \{1, \cdots, p\} : \frac{\hat{\theta}_i^2}{\hat{\Sigma}_{ii}} \sim \chi^2(1) \qquad \text{sous l'hypothèse } \theta_i = 0,$$

$$\forall i \in \{1, \cdots, p\}, \forall j \in \{1, \cdots, p\} : \frac{\left(\hat{\theta}_i - \hat{\theta}_j\right)^2}{\hat{\Sigma}_{ii} + \hat{\Sigma}_{jj} - 2\hat{\Sigma}_{ij}} \sim \chi^2(1) \qquad \text{sous l'hypothèse } \theta_i = \theta_j.$$

Ainsi le test de significativité des paramètres permettra :

– de tester les paramètres $\beta = (\beta_0', \beta_1')'$ par rapport à 0, cela revient à tester la pertinence de l'hypothèse nulle (H_0) : le facteur a une influence dans le modèle,
– de tester α_2 par rapport à α_1, ce qui consiste à tester la pertinence de distinguer les états E1 et E2.

Résultats du modèle GompitZ

Le Tableau 4.5 résume pour chacun des matériaux modélisés les paramètres significatifs (p-value < 5%).

TABLE 4.5 – Paramètres significatifs lors du calage du modèle GompitZ.

	Canalisation			Branchement	
	Fonte Grise / Amiante Ciment	Fonte Ductile / Acier	PVC/PE	PE Noir	PE Bleu
Longueur	×	×	×		
Diamètre	×	×	×		
Nombre casses historiques	×	×		×	×
Période de pose	×				×
Pression	×	×		×	×
Sol Argileux				×	×
Lotissement					×
Chaud				×	×
Froid	×				

Les résultats présentés à la Table 4.6 sont issus du calage du modèle pour le matériau Fonte Grise/Amiante Ciment (FG-AM). Les résultats pour le reste des matériaux sont présentés en Annexe H.

TABLE 4.6 – Estimation des paramètres du modèle GompitZ pour le matériau FG-AM

Facteur	$\hat{\theta}$	$\sqrt{\hat{\Sigma}_{ii}}$	p-value
α_1	-4.300	0.043	-
α_2	-4.350	0.044	$< 1 \times 10^{-5}$
Diamètre	-0.004	0.000	$< 1 \times 10^{-5}$
Intercept	-1.506	0.080	$< 1 \times 10^{-5}$
Longueur	0.502	0.011	$< 1 \times 10^{-5}$
Période Pose $\leq$ 1933	-0.863	0.025	$< 1 \times 10^{-5}$
Variation de pression	0.321	0.024	$< 1 \times 10^{-5}$
Climat (Froid)	0.045	0.002	$< 1 \times 10^{-5}$

A partir de la formule 4.6, il est possible de calculer la probabilité

$$\hat{\mathbf{P}}^i_j(t) = \mathbb{P}_{\hat{\theta}}\{X_i(t) = j | \mathbf{Z_i}\}.$$

Le calcul de cette probabilité, sur chaque individu et pour chaque état, nous permet d'évaluer la qualité d'ajustement du modèle, en comparant :

– d'une part $\sum_i \mathbb{1}_{[Y_i(t)=j]}/N$: la répartition des individus par état j,

– d'autre part $\sum_i \hat{\mathbf{P}}^i_j(t)/N$: la probabilité moyenne d'être dans un état j.

Une vision macroscopique de la qualité d'ajustement, i.e. une qualité d'ajustement sur l'ensemble des ouvrages, renvoie les résultats présentés à la Figure 4.10.

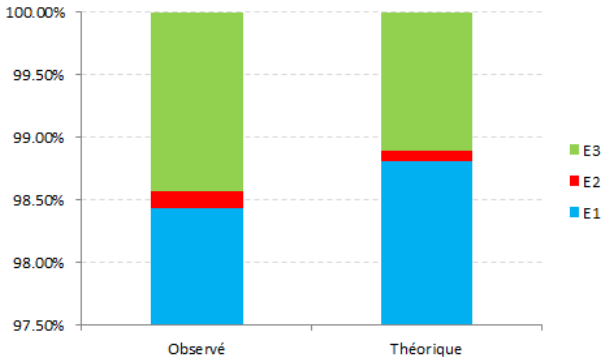


FIGURE 4.10 – Résultats d'estimation des probabilités d'appartenance à un état du modèle GompitZ

Il est possible de conclure à une bonne performance du modèle d'après cette vision macroscopique, cependant, si on détaille la qualité d'ajustement en fonction du matériau et par classe d'âge, on constate les mauvaises performances du modèle à estimer l'état de dégradation, comme le montre l'exemple en Figure 4.11.

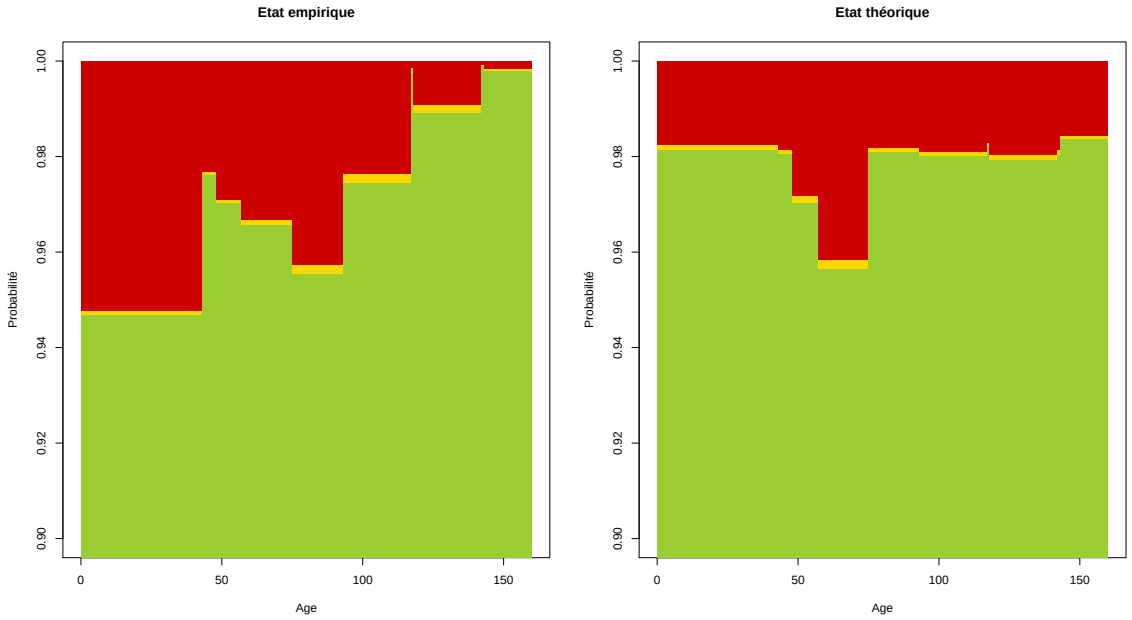


FIGURE 4.11 – Résultats d'estimation des probabilités d'appartenance à un état du modèle GompitZ pour le matériau Fonte Grise-Amiante Ciment par classe d'âge (vert = E_1, jaune = E_2, rouge = E_3)

Ainsi le modèle n'est pas précis pour l'estimation des états par ouvrage, ce modèle n'est pas adapté à notre cas d'application. Nous ne nous attarderons par sur les performances du modèle GompitZ puisqu'il ne répond pas à nos attentes : il ne modélise pas la dynamique temporelle de la dégradation mais juste la probabilité d'être dans un état à un temps donné. Nous nous tournons alors vers d'autres types de modèles : des modèles multi-états.

4.5.3 Les modèles multi-états

Le processus que nous tentons de mettre en place doit pouvoir modéliser l'évolution de la dégradation des ouvrages selon les trois états précédemment identifiés (*cf.* Figure 4.7). Cependant il faut déplorer le manque d'information apportée par les inspections. En effet, les ouvrages observés dans l'état E_2 (ouvrage subissant des fuites détectables non reportées) sont uniquement ceux qui sont repérés dans le cadre d'une RAF, et qui sont immédiatement réparés. De fait, il est impossible d'observer une transition entre l'état E_2 et l'état E_3 (ouvrage subissant des casses manifestes).

Nous définissons alors un quatrième état, l'état $E_{2'}$: il s'agit des tronçons se trouvant dans l'état E_2 et qui sont réparés suite à une RAF. Ainsi, nous introduisons un nouveau concept, celui des risques compétitifs : les tronçons se situant dans l'état E_2 peuvent donc soit être réparés suite à une campagne de RAF et donc partir dans l'état $E_{2'}$, soit se dégrader et partir dans l'état E_3. Pour évaluer l'efficacité d'ajouter un quatrième état, nous construirons différents modèles, en faisant évoluer le nombre d'états considérés.

Remarque : A l'instar des ouvrages dans l'état E_3, les ouvrages dans l'état $E_{2'}$ sont identifiés exhaustivement.

Le modèle de casse

Le premier modèle que nous considérons est le modèle le plus simple : un modèle à deux états. Nous ne considérons qu'un seul type de phénomène appelé *la défaillance de l'ouvrage*. Nous considérons qu'une défaillance est survenue si l'ouvrage a été réparé.

En d'autres termes, nous fusionnons d'une part les états E_1 et E_2 et d'autre part les états $E_{2'}$ et E_3. Ainsi, le modèle mis en place sera un modèle de type AFT (Accelerated Failure Time). De tels modèles ont déjà été appliqués au réseau d'eau potable (voir par exemple [Eisenbeis, 1994], [Røstum, 2000]) ; le modèle sélectionné est le modèle LEYP (Linear Extended Yule Process, [Le Gat, 2009]). Ce modèle tire ses origines des modèles de comptage et des évènements récurrents (voir [Babykina, 2011], chapitre 2). Le modèle LEYP est actuellement utilisé par les gestionnaires d'eau pour la prédiction des casses sur canalisations [Claudio et al., 2014].

a - Intensité du modèle LEYP pour des processus de comptage d'événements récurrents

Notons $N(t)$ le processus de comptage d'événements récurrents, l'intensité de ce processus est définit par :

FIGURE 4.12 – Le modèle LEYP : un modèle à deux états

$$\lambda(t) = \mathbb{E}[dN(t)|\mathscr{F}_{t-}] = \mathbb{P}\{dN(t) = 1|\mathscr{F}_{t-}\} \tag{4.11}$$

où $\mathscr{F}_{t-}$ est la filtration générée par le processus $N(t)$. Dans le cadre d'un LEYP, nous définissons ce processus $N(t)$ comme le processus de comptage du nombre d'occurrences de passages de E_1/E_2 vers $E_{2'}/E_3$. Le processus admet donc un retour vers les états E_1/E_2 après réparation, mais nous prenons en compte le fait que la dégradation de l'ouvrage s'intensifie avec le nombre de casses passées. Dans ce contexte, l'intensité du processus est donnée par :

$$\lambda(t,Z) = (1+\alpha N(t-))\delta t^{\delta-1}e^{Z'\beta}. \tag{4.12}$$

où $N(t-)$ est le nombre de défaillances survenues juste avant t, Z est un vecteur de covariables et $\theta = (\alpha,\delta,\beta)'$ est le vecteur de paramètres associés au modèle.

Cette intensité se décompose en trois facteurs :
- un facteur Yule ($\alpha(1+N(t-))$) qui permet de prendre en compte l'effet des défaillances passées,
- un facteur Weibull ($\delta t^{\delta-1}$) qui permet de prendre en compte l'effet du vieillissement,
- un facteur Cox ($e^{Z'\beta}$) qui permet de prendre en compte l'effet de variables explicatives.

b - Vraisemblance du modèle LEYP

Considérons un ouvrage i ($i \in [1,N]$) que l'on inspecte dans les états E2 ou E3 m_i fois aux instants t_{ij} ($j \in [1,m_i]$), observé entre les instants a_i et b_i. Sa contribution à la vraisemblance vaut [Babykina, 2011] :

$$\mathscr{L}_i(\theta) = \alpha^{m_i}\frac{\Gamma(\alpha^{-1}+m_i)}{\Gamma(\alpha^{-1})}\frac{\prod_{j=1}^{m_i}e^{\alpha\Lambda_0(t_{ij})}\lambda_0(t_{ij})}{\left(e^{\alpha\Lambda_0(b_i)}-e^{\alpha\Lambda_0(a_i)}+1\right)^{\alpha^{-1}+m_i}}. \tag{4.13}$$

où $\lambda_0(t)$ est l'intensité initial du processus définie par $\lambda_0(t) = \delta t^{\delta-1}e^{Z'\beta}$.

Le calcul de la vraisemblance peut poser problème notamment le calcul de l'intensité de base cumulée $\Lambda_0(t)$ où

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du. \tag{4.14}$$

En effet, cette intégrale n'est pas arithmétiquement soluble, notamment si le vecteur de covariables Z intègre une covariable dépendante du temps. [Babykina, 2011] propose une

solution pour pallier ce problème en transformant l'intégrale entre 0 et t en une somme d'intervalles où la covariable dépendante du temps est constante :

$$\Lambda_0(t) = \sum_{r=1}^{R} \int_{\tau_{r-1}}^{\tau_r} \delta u^{\delta-1} e^{Z(u)'\beta} du,$$

où $\tau_0 = 0$, $\tau_R = t$ et $Z(t) = Zr, \forall t \in [\tau_r, \tau_{r+1}]$.

La démarche pour estimer les paramètres est similaire à celle utilisée pour l'estimation des paramètres du modèle GompitZ. Nous calculons dans un premier temps la log-vraisemblance $l(.)$:

$$l(\theta) = \ln \mathscr{L}(\theta) = \sum_{i=1}^{N} \ln \mathscr{L}_i(\theta). \tag{4.15}$$

Nous cherchons alors le vecteur de paramètres $\hat{\theta}$ solution de l'équation (4.10). La significativité des effets des covariables est testée en utilisant un test du khi-deux de Wald pour le vecteur β (voir chapitre 4.5.2) ou un test de rapport des vraisemblances pour les paramètres α et δ (pour plus de détail, voir [Claudio et al., 2014]).

c - Résultat et validation du modèle LEYP

Un exemple d'estimation est détaillé à la Table 4.7 pour le matériau PE Bleu. Les résultats pour le reste des matériaux sont présentés en Annexe H.

TABLE 4.7 – Estimation des paramètres du modèle LEYP pour le matériau PE Bleu

Facteur	$\hat{\theta}$	$\sqrt{\hat{\Sigma}_{ii}}$	p-value
Casses historiques	4.19	0.21	$< 1 \times 10^{-5}$
Vieillissement	1.35	0.02	$< 1 \times 10^{-5}$
Intercept	-0.15	0.04	2×10^{-4}
Collier	0.27	0.05	$< 1 \times 10^{-5}$
Sol Argileux	0.23	0.03	$< 1 \times 10^{-5}$
Climat (Chaud)	0.06	0.00	$< 1 \times 10^{-5}$
Modulation de pression	-0.13	0.04	3.2×10^{-3}

L'estimation des paramètres par maximum de vraisemblance permet de calculer par la suite la probabilité qu'un ouvrage subisse au moins une casse sur une période $[a;b]$:

$$\mathbb{P}_{\hat{\theta}}\{N(b) - N(a) > 0\} = 1 - \left(\frac{1}{e^{\alpha\Lambda(b)} - e^{\alpha\Lambda(b)} + 1} \right)^{1/\alpha} \tag{4.16}$$

Ainsi nous pouvons construire la Figure 4.13 ci-après qui représente la qualité d'ajustement du modèle : pour les x premiers pourcents ayant la plus forte probabilité de casse, combien ont réellement subi au moins une casse.

On constate ainsi que le modèle LEYP est performant pour l'identification d'ouvrage situés dans les états $E_{2'}$ et E_3 : on remarque par exemple qu'en considérant les premiers 20% des ouvrages les plus à risques au sens du modèle, nous arrivons à cibler plus de 40%

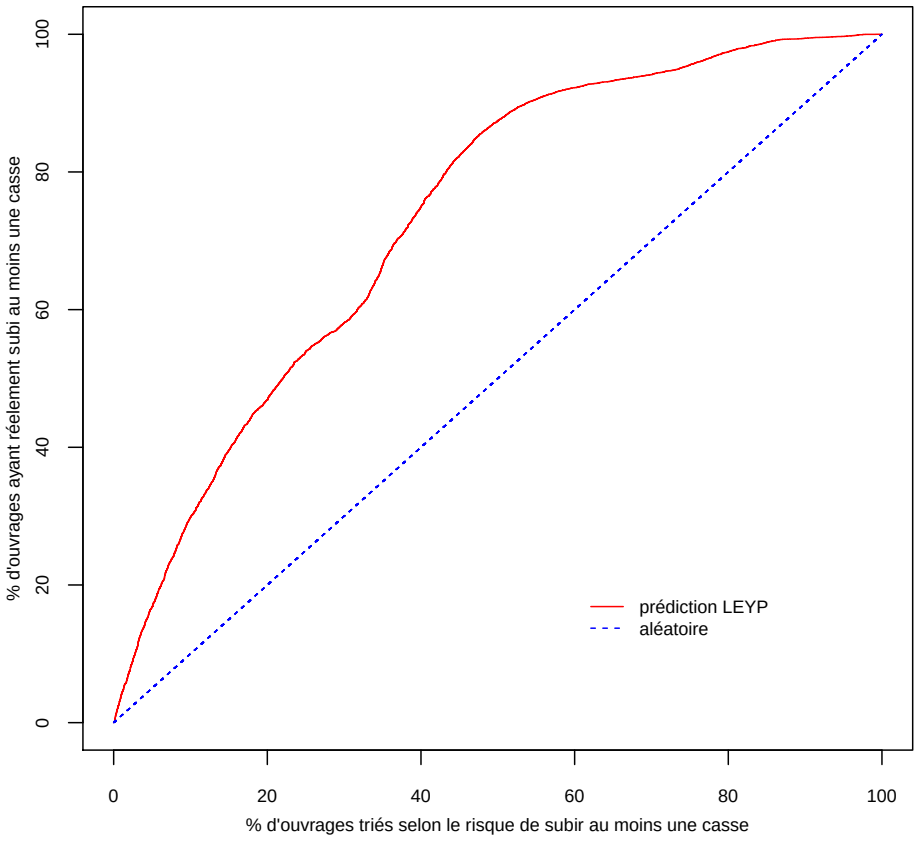


FIGURE 4.13 – Qualité d'ajustement pour le modèle LEYP

des ouvrages ayant réellement été dans les états $E_{2'}$ ou E_3. Par transposition, nous pouvons alors identifier les ouvrages situés dans les états E_1 ou E_2. Cependant, nous trouvons ici une des limites de ce modèle, dans la mesure où si ce modèle est efficace pour anticiper le nombre de casses sur le réseau, qu'elle est sa plus-value pour une réduction des fuites en "temps réel" ?

Le modèle de risques concurrents

Complexifions le modèle précédant : supposons maintenant que si un individu est réparé, cela peut être suite à une casse visible en surface ou consécutif à une recherche de fuite. Ainsi, un individu "sain" peut subir deux types j de défaillance :

– $j = 1$: l'individu a subi une casse manifeste,
– $j = 2$: l'individu subit des fuites invisibles détectées suite à une RAF.

Le modèle présenté en Figure 4.14 est quelque peu simplifié par rapport à la réalité. En effet après une réparation, un individu retourne, en quelque sorte, dans l'état E_1 et peut éventuellement subir de nouveau des défaillances. Ainsi pour disposer d'un graphe unidirectionnel, nous supposons que les états $E_{2'}$ et E_3 sont des états terminaux et que l'individu "meurt", remplacé par un nouvel individu, situé dans l'état E1, avec les mêmes

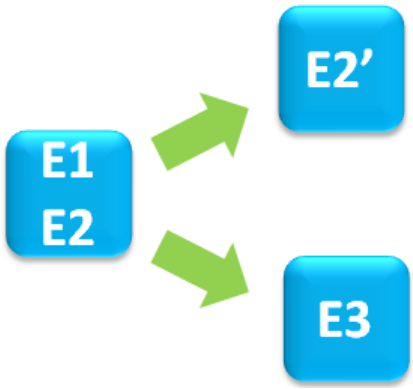


FIGURE 4.14 – Le modèle de risques compétitifs

caractéristiques intrinsèques que son prédécesseur (âge, matériau, longueur, etc.) mais avec une variable, appelée *nombre de casses historiques*, incrémentée d'une unité.

a - Intensité du modèle de risques concurrents

Soit T une variable aléatoire continue représentant le temps de survie dans l'état E_1, *i.e.* le temps T est le moment aléatoire de passage de l'état $E_{2'}$ ou E_3. Nous supposons que si une défaillance survient, elle est d'un type j ($j \in \{1,2\}$) et nous appelons J la variable aléatoire représentant le type de défaillance. Enfin, nous appelons Z un vecteur de covariables explicatives de la défaillance d'un ouvrage.

Le risque instantané global de défaillance, noté $\lambda(t,Z)$, appelé aussi fonction de hasard, est défini par (voir par exemple [Kalbfleisch and Prentice, 2002]) :

$$\lambda(t,Z) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\{t \leq T < t+\Delta t | T \geq t, Z\}}{\Delta t} \tag{4.17}$$

Ce risque instantané global correspond en réalité à la somme des risques de subir une défaillance de type j :

$$\lambda(t,Z) = \sum_{j=1}^{2} \lambda_j(t,Z), \tag{4.18}$$

où $\lambda_j(t,Z)$ représente la fonction de hasard spécifique :

$$\lambda_j(t,Z) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\{t \leq T < t+\Delta t, J = j | T \geq t, Z\}}{\Delta t} \tag{4.19}$$

Remarque : L'équation (4.19) vérifie l'égalité (4.18) par la loi des probabilités totales. En effet, nous considérons qu'une défaillance n'est due qu'à une et une seule cause.

Nous pouvons définir usuellement la fonction de survie $S(t,Z)$ (fonction de survie à tous types de risque) :

$$S(t,Z) = e^{-\Lambda(t,Z)} \tag{4.20}$$

où $\Lambda(t,Z)$ est la fonction de risque cumulée, définie par :

$$\Lambda(t,Z) = \int_0^t \lambda(u,Z)du. \tag{4.21}$$

Par analogie à $S(t,Z)$, nous pouvons définir la fonction

$$S_j(t,Z) = e^{-\Lambda_j(t,Z)} \tag{4.22}$$

où $\Lambda_j(t,Z)$ est le risque cumulé pour le type j :

$$\Lambda_j(t,Z) = \int_0^t \lambda_j(u,Z)du. \tag{4.23}$$

Notons que, en général, les fonctions S_j ne peuvent être interprétées comme des fonctions de survie, pour un nombre de causes supérieur à 2.

Remarque : A partir des équations (4.18) et (4.20), nous obtenons l'égalité :

$$S(t,Z) = \prod_{j=1}^{m} S_j(t,Z) \tag{4.24}$$

Preuve : D'après l'équation (4.20), nous avons :

$$\begin{aligned} S(t,Z) = \exp\left(-\Lambda(t,Z)\right) = \exp\left(-\int_0^t \lambda(u,Z)du\right) &= \exp\left(-\int_0^t \sum_{j=1}^{2} \lambda_j(u,Z)du\right) \\ &= \exp\left(-\sum_{j=1}^{2} \Lambda_j(t,Z)\right) \\ &= \prod_{j=1}^{2} \exp\left(-\Lambda_j(t,Z)\right) \\ &= \prod_{j=1}^{m} S_j(t,Z) \end{aligned}$$

Dans notre application, la fonction de hasard spécifique définie suit un modèle à hasards proportionnels avec une intensité initiale suivant une loi de Weibull

$$\begin{aligned} \lambda(t,Z) &= \sum_{j=1}^{2} \lambda_{j0}(t) e^{Z'\beta} \\ &= \sum_{j=1}^{2} \lambda_j \delta_j (\lambda_j t)^{\delta_j - 1} e^{Z'\beta}. \end{aligned} \tag{4.25}$$

Pour un type de défaillance j considéré, le calcul de l'intensité cumulée spécifique est simple si on considère un vecteur de covariable X fixe dans le temps :

$$\begin{aligned} \Lambda_j(t,Z) &= \int_0^t \lambda_j(u,Z)du \\ &= \int_0^t \lambda_j \delta_j (\lambda_j t)^{\delta_j - 1} e^{Z'\beta} \\ &= \left(\lambda_j t\right)^{\delta_j} e^{Z'\beta}. \end{aligned}$$

Cependant, l'intégrale n'est pas arithmétiquement soluble si $Z = Z(t)$. Cependant, comme pour le modèle LEYP, nous pouvons pallier ce problème en découpant l'intervalle de

temps $[0;t]$ en sous-intervalles où la covariable dépendante du temps est constante.

b - Vraisemblance du modèle de risques concurrents

La contribution à la vraisemblance d'un ouvrage i dans le cadre d'un modèle de risques concurrents est :

$$\mathscr{L}_i(\theta) = \prod_{j=1}^{2} \lambda_j(t_i, Z_i)^{d_{ij}} e^{-\Lambda_j(t_i,Z_i)}. \tag{4.26}$$

où d_{ij} est une indicatrice qui vaut 1 si l'individu i a subi une défaillance de type j. Comme pour les autres modèles, nous calculons la log-vraisemblance $l(.)$ telle que définie à l'équation (4.15) et nous cherchons $\hat{\theta}$ solution de l'équation (4.10).

c - Résultat et validation du modèle de risques concurrents

Le tableau 4.8 présente les résultats d'estimation des paramètres du modèles de risques concurrents pour les ouvrages (canalisation) en fonte grise/amiante ciment. Les résultats pour le reste des matériaux sont présentés en Annexe H.

TABLE 4.8 – Estimation des paramètres du modèle de risques compétitifs pour le matériau FG-AM

Facteur	$\hat{\theta}$	$\sqrt{\hat{\Sigma}_{ii}}$	p-value
$\lambda_{2'}$	0.378	0.014	$< 1 \times 10^{-5}$
λ_3	0.452	0.013	$< 1 \times 10^{-5}$
$\delta_{2'}$	6.499	0.208	$< 1 \times 10^{-5}$
δ_3	4.410	0.058	$< 1 \times 10^{-5}$
Diamètre	-0.003	0.000	$< 1 \times 10^{-5}$
Longueur	0.876	0.017	$< 1 \times 10^{-5}$
Période Pose $\leq$ 1933	-2.781	0.053	$< 1 \times 10^{-5}$
Pression	0.167	0.020	$< 1 \times 10^{-5}$
Casses historiques	0.231	0.038	$< 1 \times 10^{-5}$

Analysons à présent les performances de notre modèle de risques concurrents. Sur la Figure 4.15, nous avons représentés en parallèle la répartition des états $E_{2'}$ et E_3 d'après les observations terrain (gauche) et d'après le modèle de risques concurrents.

On constate alors au vu du graphe précédent les mauvaises performances du modèle de risques concurrents notamment pour les matériaux Fonte Grise/Amiante Ciment, PE Bleu et PE Noir. Nous ne nous attarderons pas à développer un modèle de risques concurrents performant dans la mesure où nous savons que ce type de modèle n'est pas approprié à notre contexte opérationnel. En effet, l'objectif principal du modèle de dégradation est de permettre de dissocier les états E_1 et E_2 qui sont des états non visibles en surface mais, à l'inverse du premier, le second pourrait être détecté (par recherche de fuite) et les fuites réparées. Cela conduit à considérer non pas un modèle à 3 états mais un modèle à 4 états.

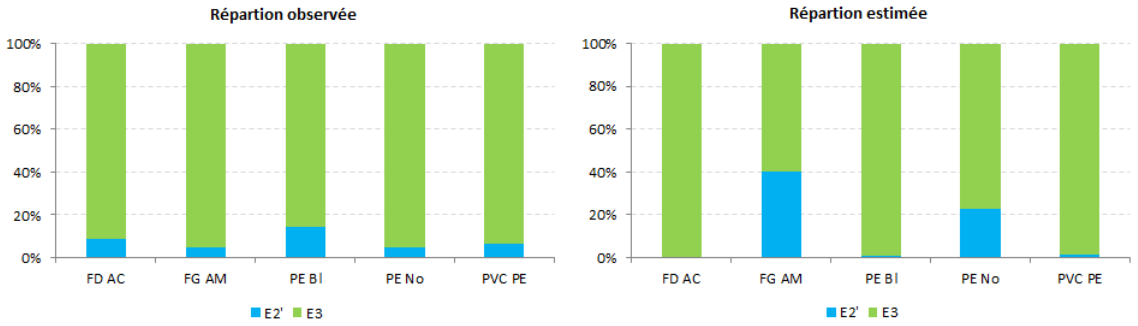


FIGURE 4.15 – Validation du modèle de risques concurrents

Le modèle semi-markovien à 4 états

Nous complexifions une nouvelle fois notre modèle : entre l'état initial et les états absorbants, nous rajoutons un état transient (état E_2) qui caractérise les ouvrages subissant des fuites invisibles mais qui peuvent être détectées lors d'une campagne de recherche de fuite. Si les ouvrages en E_2 sont identifiés et réparés lors d'une campagne de recherche de fuite, alors il transite dans l'état $E_{2'}$; si au contraire, ils ne sont pas identifiés, alors ils se dégraderont et finiront dans l'état E_3. Ce modèle est donc le plus complet mais aussi le plus complexe (voir Figure 4.16). Il caractérise les différents aspects de dégradation des fuites sur les ouvrages. Nous appellerons ce modèle par la suite le modèle "multi-états de dégradation".

FIGURE 4.16 – Le modèle multi-états de dégradation (MED)

4.5.4 Le modèle multi-états de dégradation

Définition et notation

Soit $\mathbb{X}(.) = \{\mathbb{X}(t), t \geq 0\}$ un processus temporel continu à 4 états (1,2,2',3). Nous appelons X_n la variable aléatoire désignant l'état du processus après le $n^{\text{ème}}$ saut, T_n l'instant de transition du processus au $n^{\text{ème}}$ saut (avec T_0=0) et D_n la variable aléatoire représentant la durée de séjour entre le $n-1^{\text{ème}}$ et le $n^{\text{ème}}$ saut (avec $D_n = T_n - Tn - 1$).

Le processus $\mathbb{X}(.)$ est un processus semi-markovien si :

– la sequence de variables aléatoires $\{X_0, X_1, \cdots\}$ des états occupés forme une chaîne de Markov. Les probabilités de transition d'un état i vers un état j, notées P_{ij} [1], sont notées :

$$P_{ij} = \mathbb{P}\{X_{n+1} = j | X_n = i\}. \tag{4.27}$$

– les temps de séjour D_n sont donc des variables aléatoires indépendantes dont la distribution ne dépend que de l'état de provenance :

$$\mathbb{P}\{D_{n+1} \leq d, X_{n+1} = j | X_0, D_1, X_1, \cdots, D_n, X_n\} = \mathbb{P}\{D_{n+1} < d, X_{n+1} = j | X_n\}. \tag{4.28}$$

– les intensités de transition du processus d'un état i vers un état j, notée α_{ij}, ne dépendent que de la durée de séjour dans l'état actuel et des états adjacents :

$$\alpha_{ij}(d) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}\{d < D_{n+1} < d + h, X_{n+1} = j | D_{n+1} > d, X_n = i\} \tag{4.29}$$

– nous définissons $\alpha_{ii}(d) = -\sum_{j \neq i} \alpha_{ij}(d)$ où $\sum_{j \neq i} \alpha_{ij}(d)$ est la fonction de hasard de la durée de séjour dans l'état i [Commenges and Gégout-Petit, 2007], ce qui nous permet de définir la survie dans l'état i :

$$S_i(d) = \exp\left(\int_0^d - \sum_{j \neq i} \alpha_{ij}(u) du \right). \tag{4.30}$$

Il est aussi possible de caractériser le processus semi-markovien selon :
(i) la fonction de répartition de la durée de séjour en i avant de passer en j :

$$F_{ij}(d) = \mathbb{P}\{D_{n+1} \leq d | X_{n+1} = j, X_n = i\} \tag{4.31}$$

(ii) la fonction de survie en i avant de passer en j :

$$S_{ij}(d) = 1 - F_{ij}(d) = \mathbb{P}\{D_{n+1} > d | X_{n+1} = j, X_n = i\} \tag{4.32}$$

(iii) la densité de la durée de séjour en i avant de passer en j :

$$f_{ij}(d) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}\{d < D_{n+1} < d + h | X_{n+1} = j, X_n = i\} \tag{4.33}$$

D'après [Foucher et al., 2007], il est possible de faire le lien entre les séquences d'états et les durées de séjour :

$$\alpha_{ij}(d) = \frac{P_{ij} f_{ij}(d)}{S_i(d)}. \tag{4.34}$$

1. Dans notre travail, nous ne nous intéressons qu'au transition de i vers j où i n'est pas un état absorbant. De fait par la suite, $P_{ii} = 0, \forall i$.

Encadré 2 : L'intensité de transition α_{ij}

Nous définissons les évènements suivants :

A : "$d < D_{n+1} < d+h$"
B : "$X_{n+1} = j$"
C : "$X_n = i$"
D : "$D_{n+1} > d$" $\Rightarrow A \subset D \Leftrightarrow A \cap D = A$

Nous avons alors :

$$\begin{aligned} P_{ij} &= \mathbb{P}(B|C) = \mathbb{P}(B \cap C)/\mathbb{P}(C) \\ f_{ij}(d) &= \lim_{h\to 0} \tfrac{1}{h}\, \mathbb{P}(A|B \cap C) = \lim_{h\to 0} \tfrac{1}{h}\, \mathbb{P}(A \cap B \cap C)/\mathbb{P}(B \cap C) \\ S_i &= \mathbb{P}(D|C) = \mathbb{P}(D \cap C)/\mathbb{P}(C) \end{aligned}$$

En partant du membre de droite de l'équation(4.34), nous obtenons :

$$\begin{aligned} \frac{P_{ij} f_{ij}(d)}{S_i(d)} &= \lim_{h\to 0} \frac{1}{h} \frac{\mathbb{P}(B \cap C)\, \mathbb{P}(A \cap B \cap C)\, \mathbb{P}(C)}{\mathbb{P}(C)\, \mathbb{P}(\cap B \cap C)\, \mathbb{P}(D \cap C)} \\ &= \lim_{h\to 0} \frac{1}{h} \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(D \cap C)} \\ &= \lim_{h\to 0} \frac{1}{h} \frac{\mathbb{P}(A \cap B \cap C \cap D)}{\mathbb{P}(D \cap C)} \qquad \text{car } A \subset D \\ &= \lim_{h\to 0} \frac{1}{h} \mathbb{P}(A \cap B | C \cap D) \\ &= \alpha_{ij}(d) \end{aligned}$$

D'après les formules (4.30) et (4.34) nous avons (voir [Kang and Lagakos, 2007]) :

$$P_{ij} f_{ij}(d) = \alpha_{ij}(d) S_i(d) = \alpha_{ij}(d) \exp\left(\int_0^d - \sum_{j \neq i} \alpha_{ij}(u) du \right). \tag{4.35}$$

Remarque : A partir des équations précédentes, nous pouvons définir la fonctions de répartition de durée dans l'état i, la survie dans l'état i :

$$\begin{aligned} F_i(d) &= \mathbb{P}\{D_{n+1} \leq d | X_n = i\} \\ &= \sum_j \mathbb{P}\{D_{n+1} \leq d, X_{n+1} = j | X_n = i\} \\ &= \sum_j \mathbb{P}\{X_{n+1} = j | X_n = i\} \mathbb{P}\{D_n \leq d | X_{n+1} = j, X_n = i\} \\ &= \sum_{j \neq i} P_{ij} F_{ij}(d) \qquad \text{(puisque } P_{ii} = 0\text{)}. \end{aligned} \tag{4.36}$$

Ainsi que la densité marginale :

$$\begin{aligned} f_i(d) = \partial F_i(d)/\partial d &= \partial \left(\sum_{j \neq i} P_{ij} F_{ij}(d) \right) / \partial d \\ &= \sum_{j \neq i} P_{ij} \left(\partial F_{ij}(d)/\partial d \right) \\ &= \sum_{j \neq i} P_{ij} f_{ij}. \end{aligned} \tag{4.37}$$

Remarque : D'après l'équation (4.35), nous avons

$$P_{ij} F_{ij}(d) = \int_0^d P_{ij} f_{ij}(u) du = \int_0^d \alpha_{ij}(u) \exp \left(\int_0^u - \sum_{j \neq i} \alpha_{ij}(v) dv \right) du.$$

En notant que $\lim_{d \to +\infty} F_{ij}(d) = 1$, nous avons $\lim_{d \to +\infty} P_{ij} F_{ij}(d) = P_{ij}$ et de fait nous obtenons :

$$P_{ij} = \lim_{d \to +\infty} \int_0^d \alpha_{ij}(u) \exp \left(\int_0^u - \sum_{j \neq i} \alpha_{ij}(v) dv \right) du. \tag{4.38}$$

Application à la dégradation des ouvrages d'eau potable

Si nous appliquons notre modèle à notre cas d'étude, nous pouvons affirmer d'après la Figure 4.16 que :

- les états E_1 et E_2 sont des états transients,
- les états $E_{2'}$ et E_3 sont des états terminaux,
- il n'existe que 3 transitions possibles : $E_1 \to E_2$, $E_2 \to E_{2'}$, $E_2 \to E_3$.

Ainsi, il est possible de simplifier les écritures :
- $P_{12} = 1$,
- $f_1(d) = f_{12}(d)$, $F_1(d) = F_{12}(d)$ et $S_1(d) = S_{12}(d)$.

Une des particularité du modèle que nous voulons mettre en place provient aussi des données exploitées. Les opérations réalisées sur le réseau sont l'unique source de données dont nous disposons pour attribuer à un ouvrage l'état auquel il appartient. Or il est important de souligner que ces données ne correspondent pas aux instants de saut dans chaque état mais à une inspection d'un ouvrage dans un état à un instant donné, sans vraiment savoir depuis combien de temps il y est ; nous n'avons qu'une vue partielle de sa trajectoire. Ces données sont appelées "données de panel" [Kalbfleisch and Lawless, 1985].

Plusieurs auteurs ont travaillé sur ce type d'observations ([Lagakos et al., 1978], [Kang and Lagakos, 2007], [Foucher et al., 2007]), ce schéma d'observation étant fréquemment observé en biostatistique. [Commenges and Gégout-Petit, 2007] définissent le

schéma d'observation dont nous disposons comme étant des observations mixées continues et discrètes : certains états sont observés de façon discrète (de manière irrégulière et sans lien avec l'instant de saut dans un état) alors que d'autres états (généralement des états absorbants) sont observés de façon continue, les observations correspondant aux instants de saut.

Nous pouvons faire l'hypothèse que les états $E_{2'}$ et E_3 sont enregistrés exhaustivement et que la date enregistrée dans les fichiers correspond à la date de saut dans l'état. Il faut noter qu'aucun ouvrage n'est observé dans l'état E_2, mais nous savons, d'après la définition de notre processus (voir Figure 4.16) que si un ouvrage arrive dans un des états terminaux, c'est qu'il provenait forcément de l'état E_2. La Figure 4.17 résume l'ensemble des trajectoires observables pour le modèle MED. Les "|" désignent l'observation d'un individu dans un état à un instant t_j, les "$\times$" désignent la censure à droite, c'est-à-dire la fin de l'observation d'un individu au moment T (fin de la fenêtre d'observation).

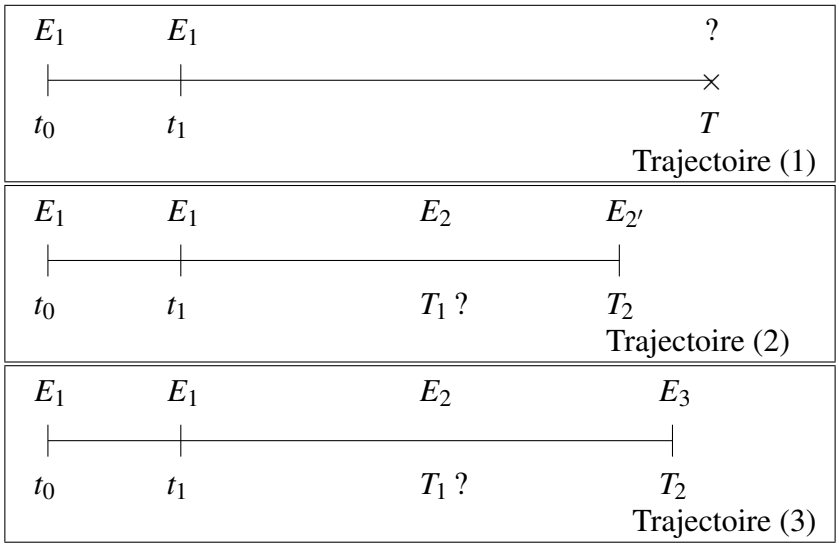


FIGURE 4.17 – Trajectoires observables pour le modèle MED

Probabilités d'appartenance à un état

L'objectif de ce modèle est de fournir les données d'entrée qui alimenteront le modèle de décomposition de débit de fuite. Ainsi les résultats attendus du modèle multi-états de dégradation sont les probabilités, pour chaque ouvrage, d'appartenir, à un instant t aux états E_1, E_2, $E_{2'}$ ou E_3.

$$P_1(t) = \mathbb{P}\{X(t) = E_1 | X_0 = E_1\} = S_1(t) = \exp\left(-\int_0^t \alpha_{12}(u)du\right), \tag{4.39}$$

$$\begin{aligned} P_2(t) = \mathbb{P}\{X(t) = E_2 | X_0 = E_1\} &= \mathbb{P}\{D_1 < t \cap D_2 > t - D_1\} \\ &= \mathbb{E}\left(\mathbb{1}[D_1 < t \cap D_2 > t - D_1]\right) \\ &= \int_0^{+\infty} \int_0^{+\infty} \mathbb{1}[d_1 < t \cap d_2 > t - d_1] f_1(d_1) f_2(d_2) \\ &= \int_0^t \left[\int_{t-d_1}^{+\infty} f_2(d_2) dd_2 \right] f_1(d_1) dd_1 \\ &= \int_0^t S_2(t - d_1) f_1(d_1) dd_1, \end{aligned} \tag{4.40}$$

$$P_{2'}(t) = \mathbb{P}\{X(t) = E_{2'} | X_0 = E_1\} = P_{22'} \mathbb{P}\{D_1 + D_2 < t\} = P_{22'} \int_0^t \int_0^v f_1(u) f_2(v-u) du dv, \tag{4.41}$$

$$P_3(t) = \mathbb{P}\{X(t) = E_3 | X_0 = E_1\} = P_{23} \mathbb{P}\{D_1 + D_2 < t\} = P_{23} \int_0^t \int_0^v f_1(u) f_2(v-u) du dv, \tag{4.42}$$

avec $P_{22'} + P_{23} = 1$.

Estimation des paramètres : contribution à la vraisemblance

Chaque individu contribue à la vraisemblance en fonction de l'information qu'il apporte, *i.e.* de sa trajectoire. [Foucher et al., 2010] identifient, en fonction de différentes trajectoires possibles, la contribution à la vraisemblance ; nous ne nous intéresserons ici qu'aux contributions liées aux trajectoires présentées à la Figure 4.17.

Trajectoire (1) : Un individu i est fixé au moment de sa pose (t_0) dans l'état E_1, puis a été observé de nouveau dans cet état E_1 à un instant t_1. Aucune autre observation n'a été faite sur cet individu entre t_1 et la fin de la fenêtre d'observation T. Nous savons que cet individu n'est pas allé dans un des états terminaux E_K ($E_{2'}$ ou E_3) car l'entrée dans un de ces états implique une réparation et donc une observation. Ainsi, entre les instants t_1 et T, l'individu est soit resté dans l'état E_1, soit entré dans l'état E_2, à un instant inconnu, sans y être sorti à T. La contribution à la vraisemblance de cet individu, notée $C_{i,1}$ s'écrit :

$$\begin{aligned} C_{i,1} &= S_1(T) + \int_{t_1}^T f_1(u) \sum_{j=2'}^{3} P_{2j} S_{2j}(T-u) du \\ &= S_1(T) + \int_{t_1}^T f_1(u) S_2(T-u) du. \end{aligned} \tag{4.43}$$

Trajectoire (2) : Un individu i est fixé au moment de sa pose (t_0) dans l'état E_1, puis est observé de nouveau dans E_1 à un instant t_1. Il est entré dans l'état terminal $E_{2'}$ à l'instant T_2. Nous savons donc que cet individu est entré dans l'état E_2 à un instant compris entre t_1 et T_2 et qu'il est sorti de cet état pour entrer dans un état terminal $E_{2'}$ à l'instant T_2. La contribution à la vraisemblance de cet individu, notée $C_{i,2}$ s'écrit :

$$C_{i,2} = P_{22'} \int_{t_1}^{T_2} f_1(u) f_{22'}(T_2 - u) du. \tag{4.44}$$

Trajectoire (3) : Similaire à la trajectoire (2), l'individu est observé à l'instant T_2 dans l'état E_3. La contribution à la vraisemblance de cet individu, notée $C_{i,3}$ s'écrit :

$$C_{i,3} = P_{23} \int_{t_1}^{T_2} f_1(u) f_{23}(T_2 - u) du. \tag{4.45}$$

On note d_{iq} la variable indicatrice qui vaut 1 si l'individu i a suivi la trajectoire q ($q \in \{1,2,3\}$). La vraisemblance globale s'écrit

$$\mathscr{L}(\theta) = \prod_{i=1}^{N} \prod_{q=1}^{3} {C_{i,q}}^{d_{iq}}. \tag{4.46}$$

Estimation des paramètres : résultats

D'après les équations 4.35 à 4.30, nous constatons que les différentes informations (densité, fonction de survie, etc.) s'expriment en fonction de l'intensité de transition. Nous décidons de modéliser cette dernière suivant une distribution de Weibull de paramètres (η_{ij}, δ_{ij}). Comme nous l'avons souligné précédemment (*cf.* partie 4.5.1), certains facteurs peuvent expliquer la transition vers les états terminaux, c'est pourquoi nous incluons dans l'écriture des intensités de transitions un vecteur Z de covariables explicatives. De fait l'intensité de transition d'un état i vers un état j prend la forme suivante

$$\alpha_{ij}(t) = \frac{\delta_{ij}}{\eta_{ij}} \left(\frac{t}{\eta_{ij}} \right)^{\delta_{ij}-1} \exp(Z'\beta) \tag{4.47}$$

L'estimation des paramètres par maximum de vraisemblance a permis d'obtenir les résultats présentés à la Table 4.9 (ces résultats concernent les ouvrages en fonte grise/amiante ciment, les résultats pour l'ensemble des matériaux sont présentés en annexe H).

TABLE 4.9 – Estimation des paramètres du modèle multi-états de dégradation pour le matériau FG-AM

Facteur	$\hat{\theta}$	$\sqrt{\hat{\Sigma}_{ii}}$	p-value [1]
δ_{12}	3.903	0.194	$< 1 \times 10^{-5}$
$\delta_{22'}$	0.821	0.041	$< 1 \times 10^{-5}$
δ_{23}	0.646	0.001	$< 1 \times 10^{-5}$
η_{12}	9.031	0.449	$< 1 \times 10^{-5}$
$\eta_{22'}$	717.221	35.620	$< 1 \times 10^{-5}$
η_{23}	91.223	4.531	$< 1 \times 10^{-5}$
Intercept	5.978	0.297	$< 1 \times 10^{-5}$
Diamètre	-0.001	0.000	$< 1 \times 10^{-5}$
Longueur	0.580	0.001	$< 1 \times 10^{-5}$
Période Pose	-2.355	0.020	$< 1 \times 10^{-5}$
Casses historiques	0.234	0.012	$< 1 \times 10^{-5}$

La Figure 4.18 représente les fonctions de survie pour les ouvrages en Fonte Grise/Amiante Ciment selon les paramètres présentés à la table 4.9.

1. les paramètres δ_{ij} sont testés par rapport à 1.

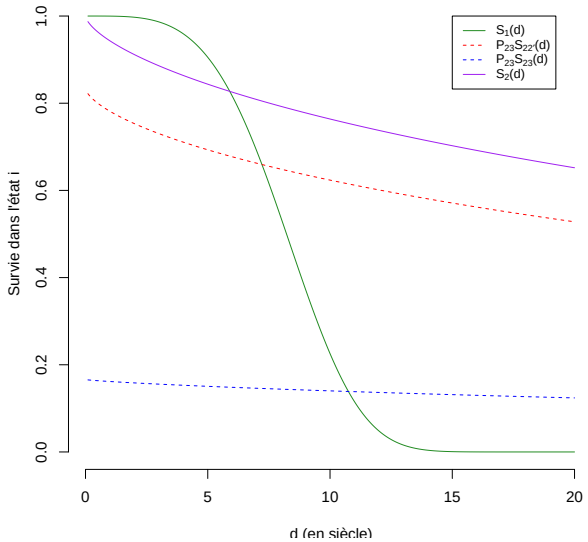


FIGURE 4.18 – Fonction de survie pour les ouvrages en Fonte Grise - Amiante Ciment

On constate à la figure 4.18 que les ouvrages en Fonte Grise/ Amiante Ciment ont tendance à rapidement transiter de l'état 1 vers l'état 2 ; alors qu'inversement, la durée de séjour dans l'état 2 avant la sortie vers un des états terminaux est plus grande. Dans ce cas, les ouvrages semblent avoir une plus grande tendance à partir vers l'état E_3. Ces résultats s'expliquent en grande partie par le fait que, comparé au nombre de casses manifestes observés, peu de fuites sont détectées par recherche de fuites sur le réseau bordelais.

Par ailleurs, les coefficients des covariables explicatives ont des valeurs proches de celles que l'on peut obtenir avec les autres modèles (type LEYP), en particulier celui lié à la longueur. Cela confirme que l'intensité de transition est proportionnelle à la racine carrée de la longueur.

Comme nous l'avons dit précédemment, l'objectif de ce modèle est de nous permettre d'estimer les probabilités d'appartenance à chaque état de dégradation. Ainsi, nous avons pu calculer pour chaque ouvrage inspecté la probabilité qu'il se situe dans un des états E_1, $E_{2'}$ et E_3 et que nous comparons aux inspections terrain selon le même procédé que celui utilisé en section 4.5.2. Ainsi la figure 4.19 représente la qualité d'ajustement du modèle.

Les résultats présentés à la figure 4.19 paraissent satisfaisants, le modèle semble être adapté à notre cas d'étude : tout d'abord il retranscrit bien l'évolution de la dégradation décrite par les experts des réseaux d'eau, il prend en compte la réalité du terrain (des fuites détectées avant l'atteinte de la casse manifeste) et s'ajuste correctement aux données d'exploitation. Ainsi, ce dernier modèle construit semble le plus abouti et le plus adapté des modèles pour venir alimenter le modèle de décomposition du débit de fuite.

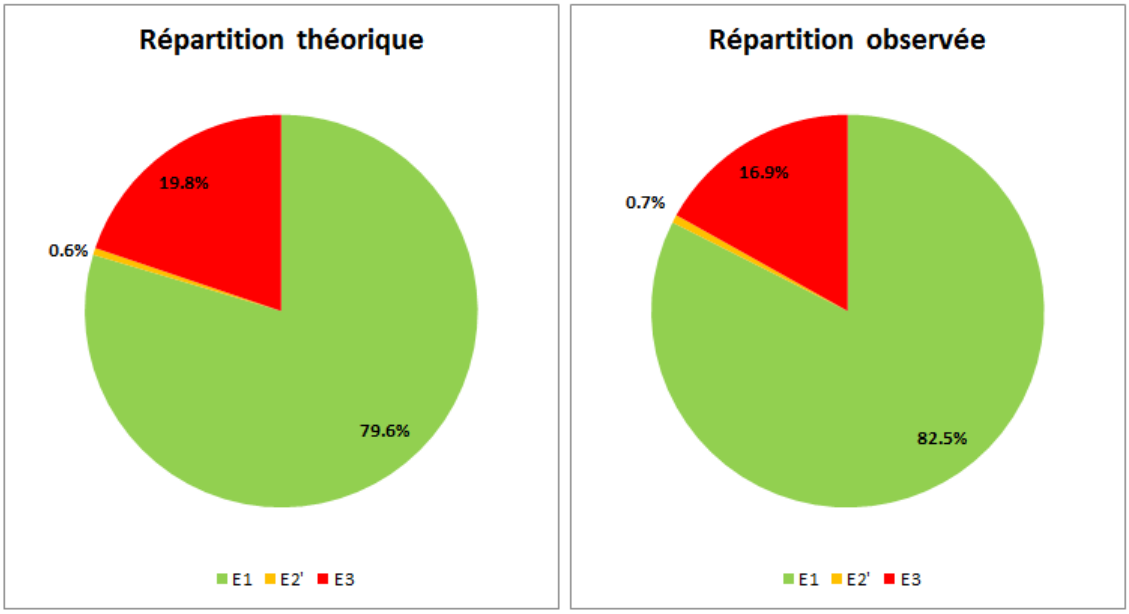


FIGURE 4.19 – Qualité d'ajustement (canalisations et branchements) du modèle multi-états de dégradation

4.6 La décomposition de la chronique de pertes selon les états de dégradation

La décomposition de la chronique de pertes est le résultat du croisement de deux modèles : le modèle multi-états de dégradation (vu à la section précédente) et le modèle de débit de fuite. Ce dernier permet de calculer le débit de pertes pour un ouvrage et pour un état de dégradation donné. Ainsi le croisement des deux modèles permet de calculer le volume d'eau perdue par état de dégradation et par ouvrage. Notons que par la suite, nous revenons à une dégradation des ouvrages selon trois états et que les débits de fuites seront modélisés pour les états $E1$, E_2+$E_{2'}$ et E_3. Il est important de distinguer dans le processus de dégradation les états E_2 et $E_{2'}$ mais cette distinction n'a plus d'intérêt lorsque l'on s'intéresse aux volumes perdus.

4.6.1 Le modèle de débit de fuite

Le modèle de débit de fuite permet de calculer le volume d'eau perdu par un ouvrage. Nous l'avons vu précédemment, la variation du débit de fuite est dépendant de la variation de la pression ; de fait, le modèle de débit de fuite intègre des données de pression. [Van Zyl and Clayton, 2007] rapportent des résultats du *Water Research Group* de l'université de Johannesbourg démontrant l'impact du diamètre de l'ouvrage dans le débit de fuite ; le modèle de débit de fuite dépend aussi du diamètre. Enfin, d'après la théorie du concept BABE [Lambert, 1994], nous savons que le débit d'une fuite dépend de l'état de dégradation de l'ouvrage (voir Figure 1.1). Ainsi, le modèle de débit de fuite dépend de la pression, du diamètre de l'ouvrage et de l'état dans lequel il se trouve. Notons $Q_j(t, \mathbf{Y_i}, \Theta)$ le débit moyen de fuite au jour t d'un ouvrage i situé dans un état j, avec un vecteur de

covariables $\mathbf{Y_i} = (Pr_i, D_i)'$:

$$Q_j(t, \mathbf{Y_i}, \Theta) = \phi_j Pr_i(t)^{\nu_k} D_i^{\gamma}. \tag{4.48}$$

où $Pr(t)$ désigne la pression au réseau au temps t et D désigne le diamètre de l'ouvrage (pour les branchements, nous considérerons que le diamètre vaut 1 ce qui n'aura pas d'incidence sur le modèle).

Ici le vecteur de paramètres est $\Theta = (\phi, \nu, \gamma)'$ où $\phi = (\phi_1, \phi_2, \phi_3)'$ dépend de l'état de dégradation et $\nu = (\nu_1, \nu_2, \nu_3, \nu_4, \nu_5)'$ dépend du matériau. Le paramètre ν_k joue un rôle similaire à l'exposant N_1 qui relie la variation de pression à la variation du débit de fuite (cf. équation (2.1), chapitre 2).

Pour caler ce deuxième modèle, nous exploitons les données de pertes estimées calculées à partir de l'estimation des consommation (voir par exemple la Figure 4.2). Le principe est d'utiliser d'une part $\mathbf{P}_j(t)$ les probabilités d'appartenance à un état de dégradation (obtenues grâce au modèle multi-états de dégradation) mais aussi du débit de fuite associé à cet état $Q_j(t, \mathbf{Y_i})$ pour calculer un volume de pertes en eau journalières $\tilde{L}(t)$ comme le montre l'équation (4.49) :

$$\tilde{L}(t, \Theta) = \sum_{j=1}^{3} \tilde{L}_j(t, \Theta) = \sum_{j=1}^{3} \sum_{i=1}^{N} Q_j(t, \mathbf{Y_i}, \Theta) \times \mathbf{P}_j(t, \mathbf{Z_i}, \theta). \tag{4.49}$$

Le calage de ce modèle consiste alors à trouver le jeu de paramètres $\hat{\hat{\theta}}$ qui permettent d'approcher au mieux des pertes $\hat{L}(t)$ (estimation par moindres carrés) :

$$\hat{\Theta} = \arg\min_{\theta} \{ \sum_{t} (\hat{L}(t) - \tilde{L}(t, \theta))^2 \} \tag{4.50}$$

4.6.2 Application du modèle au réseau de la CUB

Les deux modèles à mettre en place ont tous les deux des pré-requis :
- le modèle multi-états nécessite une connaissance exhaustive du patrimoine (a minima géo-référencement et date de pose),
- le modèle de débit de fuite nécessite de connaître les volumes d'eau perdue.

Afin de répondre au premier critère, il est nécessaire de s'intéresser aux données de la Communauté Urbaine de Bordeaux et pour satisfaire le second pré-requis, il faut alors se focaliser sur le sous-secteur du Bas Cenon, appelé aussi "Cote 50" (voir carte 2.10 à la fin du chapitre 2). Ce secteur est en effet le seul sur la CUB à disposer d'un échantillon de compteurs (composé d'usagers domestiques et industriels) équipés en émetteurs télé-relevé.

Le secteur du Bas Cenon ne représente que 0.5% du réseau de la CUB et ne peut être qualifié de représentatif en termes de répartition des matériaux comme le montre les figures 4.20 et 4.21.

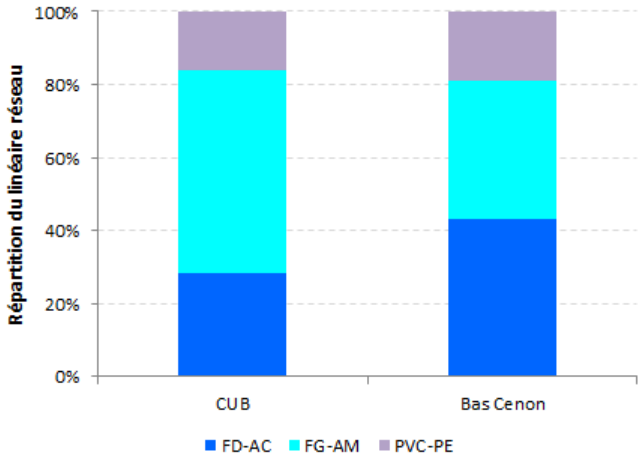


FIGURE 4.20 – Comparaison du patrimoine canalisations entre le réseau de la CUB et le sous-secteur du Bas Cenon

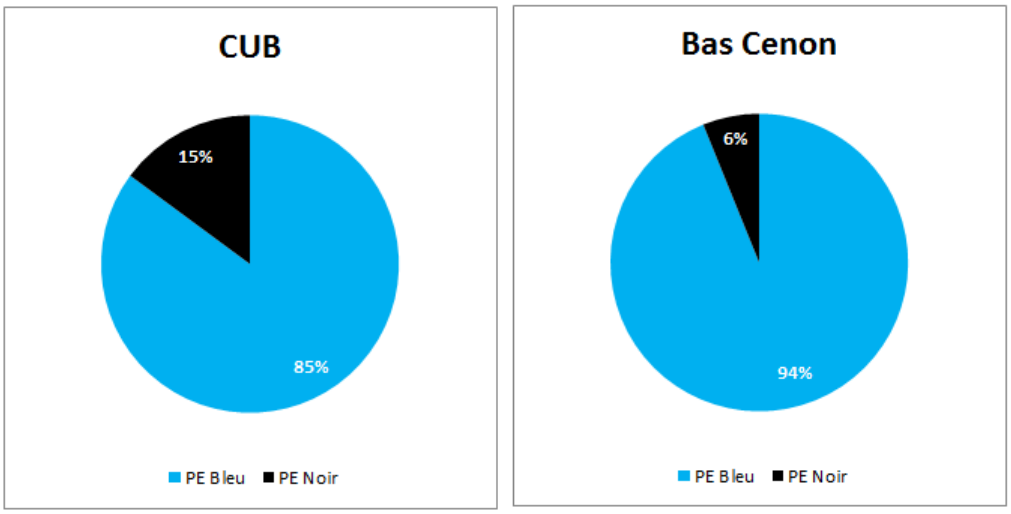


FIGURE 4.21 – Comparaison du patrimoine branchements entre le réseau de la CUB et le sous-secteur du Bas Cenon

Enfin on peut noter que la répartition canalisations/branchements n'est pas non plus respectées car on compte sur la CUB une moyenne de 40 branchements modélisés par kilomètre de canalisation modélisées contre 60 pour le Bas Cenon.

Même si la répartition des matériaux ou des ouvrages entre les deux secteurs n'est pas respectée, il n'est pas possible d'extraire un échantillon d'ouvrages sur le Bas Cenon représentatif du réseau de la CUB car la décomposition des pertes doit se calculer sur l'intégralité du patrimoine d'un secteur.

Pour caler ce modèle, nous nous basons sur les données de pertes estimées entre le 01/01/2012 et le 10/05/2013. La Figure 4.22 représente les pertes journalières qu'il faudra ajuster à l'aide du modèle de décomposition de débit de fuite.

Comme nous l'avons vu précédemment (*cf.* section 4.1) la chronique de pertes en eau est fortement bruitée. Dans ce cas précis, le bruit est principalement dû au faible taux

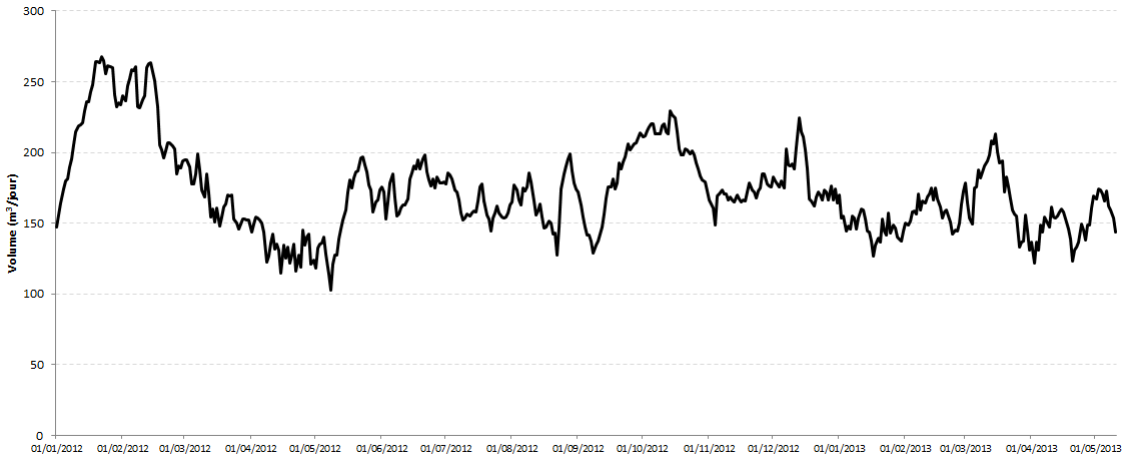


FIGURE 4.22 – Chronique des pertes à décomposer à l'aide du modèle de débit de fuite

d'échantillonnage (seulement 10% de la population). Sur la chronique étudié, l'écart-type de la consommation totale estimée représente en moyenne 8% du volume estimé. Il faut mettre en avant cette forte imprécision sur les pertes qui peut expliquer les potentiels mauvais résultats du modèle de décomposition, dans la mesure où l'évolution des fuites n'est pas due à la dégradation, ni à l'évolution de la pression, mais au bruit généré par l'estimation.

4.6.3 Résultats du modèle

Le calage du modèle de débit de fuite sur les données présentées en Figure 4.22 a permis d'obtenir les résultats présentés à la table 4.10.

TABLE 4.10 – Estimation des paramètres du modèle de débit de fuite

Facteur	$\hat{\Theta}$	$\sqrt{\hat{\Sigma}_{ii}}$	p-value
ϕ_1	0.0005	0.002	$H_0 : \phi_1 = 0$, p-value = 0.84
ϕ_2	0.0021	0.003	$H_0 : \phi_2 = \phi_1$, p-value = 0.13
ϕ_3	0.1299	0.002	$H_0 : \phi_3 = \phi_2$, p-value $< 1 \times 10^{-5}$
$\nu_{\text{FD-AC}}$	0.2462	0.007	$H_0 : \nu_{\text{FD-AC}} = 0$, p-value $< 1 \times 10^{-5}$
$\nu_{\text{FG-AM}}$	0.2311	0.006	$H_0 : \nu_{\text{FG-AM}} = 0$, p-value $< 1 \times 10^{-5}$
$\nu_{\text{PE Bleu}}$	0.1799	0.001	$H_0 : \nu_{\text{PE Bleu}} = 0$, p-value $< 1 \times 10^{-5}$
$\nu_{\text{PE Noir}}$	0.3323	0.002	$H_0 : \nu_{\text{PE Noir}} = 0$, p-value $< 1 \times 10^{-5}$
$\nu_{\text{PVC-PE}}$	0.3541	0.011	$H_0 : \nu_{\text{PVC-PE}} = 0$, p-value $< 1 \times 10^{-5}$
γ	0.1604	0.008	$H_0 : \gamma = 0$, p-value = $< 1 \times 10^{-5}$

L'application de la formule 4.49 permet d'estimer le volume de pertes par état ($\tilde{L}_j(t, \hat{\Theta})$), permettant ainsi de générer la figure 4.23 ci-après.

Comme nous l'avions annoncé précédemment, la décomposition ne s'ajuste pas parfaitement aux pertes estimées (dû au bruit présent sur la chronique de pertes). La table 4.10 ainsi que la figure 4.23 permettent toutefois de tirer les conclusions suivantes :

– le coefficient ϕ est d'autant plus important que l'état de dégradation est avancé ; cela concorde avec la réalité physique, le débit d'une fuite augmente avec la dégradation,

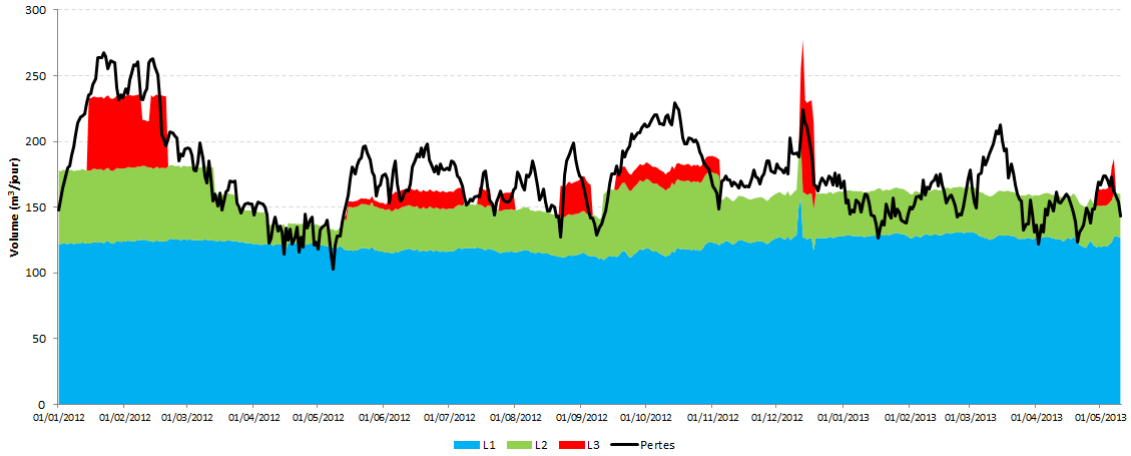


FIGURE 4.23 – Décomposition de la chronique des pertes selon les trois états de dégradation

– à l'inverse, le volume d'eau perdue est d'autant plus faible que l'état de dégradation est important ; cela s'explique par le fait que des fuites dans un état de dégradation avancé sont plus facilement détectables et donc plus rapidement réparables,
– les coefficients ν sont tous du même ordre de grandeur (entre 0.17 et 0.39), il n'y pas de distinction notable entre les matériaux plastiques et métalliques.

Au vu des coefficients ν, nous remarquons la différence avec le coefficient N_1 décrit dans la littérature. Cette différence peut s'expliquer d'une part par la répartition particulière du patrimoine ; mais elle s'explique aussi par la "mauvaise" estimation des pertes sur le réseau (ou en tout cas l'estimation imprécise des pertes).

Enfin un dernier résultat exploitable est la répartition moyenne des pertes sur la période considérée, selon les trois états de dégradation, tel que présentée à la figure 4.24.

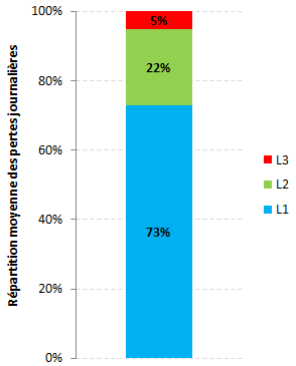


FIGURE 4.24 – Répartition moyenne des pertes selon les états de dégradation

On constate ainsi que 75% des pertes sur le secteur sont dues à des fuites invisibles et indétectables. La recherche de fuites n'est donc pas prioritaire sur ce secteur et on comprend pourquoi la modulation de pression a été mise en place en 2014 : la réduction de la pression la nuit permet de réduire le débit des fuites (tous types de fuites) et ainsi réduire les pertes même celles liées aux fuites invisibles et indétectables.

4.7 Conclusion du chapitre

Durant ce chapitre, nous nous sommes intéressés à la décomposition du volume de pertes selon trois états de dégradation des ouvrages de distribution d'eau potable. Ces états de dégradation permettent de définir et orienter les opérations à mettre en œuvre afin de lutter contre les pertes sur le réseau. Le modèle que nous avons construit pour y parvenir est le résultat de l'union de deux "sous-modèles" : le premier permet d'estimer la probabilité pour un ouvrage de se trouver dans un état de dégradation et le second permet d'évaluer le débit de fuite associé à un ouvrage et un état de dégradation donné.

Le modèle multi-états de dégradation nous permet d'estimer la probabilité pour un ouvrage de se situer dans un des états du BABE concept. L'utilisation d'un modèle semi-markovien a permis de modéliser la transition entre les différents états ainsi que les durées de séjour. Le modèle que nous avons sélectionné semble être, au vu de la comparaison des différents modèles construits, le plus complet et le plus adapté au contexte opérationnel. Les différents résultats obtenus montrent que les ouvrages ont tendance à plus rapidement transiter en dehors de l'état E_1 que de transiter en dehors de l'état E_2. Cependant, encore une fois ces informations sont à relativiser ; le manque d'information que nous avons sur les fuites invisibles permet difficilement de disposer de données exhaustives permettant un calage du modèle en adéquation avec la littérature et la " réalité terrain".

Le modèle de débit de fuite nous permet de décomposer le volume de pertes selon les trois états du concept BABE. D'après les résultats du modèle, le débit de fuite est d'autant plus important que la dégradation est importante et inversement le volume de pertes lié à un état de dégradation est d'autant plus fort que la dégradation est faible. Ce phénomène concorde avec la réalité du terrain dans la mesure où les fuites liées à un état de dégradation important ont un fort débit, sont donc plus facilement détectables et ont donc des durées d'écoulement plus faibles. La décomposition de la chronique de pertes permet en particulier de distinguer dans le volume de pertes globales, quelle est la part liée à des fuites indétectables et quelle part est liée à des fuites détectables. La comparaison de la décomposition de la chronique sur différents secteurs permettrait de hiérarchiser les actions à mettre en œuvre pour lutter contre les pertes (renouvellement, recherche de fuites, etc.).

Cependant, nous pouvons trouver des limites aux modèles présentés dans ce chapitre. L'une des premières critiques que l'on peut émettre sur les modèles concerne la représentativité des individus (ouvrages) ayant permis de caler nos modèles. Pour le premier modèle (modèle multi-états), même s'il existe dans la littérature des modèles permettant de traiter des données de panel, il faut mettre en avant deux faiblesses liées à nos données :

- nous ne disposons d'aucune donnée d'ouvrages situés dans l'état E_2,
- nous ne disposons que d'un échantillon de tronçons dans l'état E_1.

Concernant ce dernier point, on peut remarquer que l'échantillon dont nous disposons est à la fois temporel (nous n'avons que quelques inspections dans l'état E_1 pour un tronçon, lors des périodes de recherche de fuite) mais il s'agit aussi d'un échantillon spatial (les données ne sont disponibles que pour les ouvrages se situant dans des zones où il y

a une recherche de fuite). Le fait de disposer d'un échantillon temporel ne pose pas de souci particulier, cela correspond aux caractéristiques des données de panel ; le problème majeur vient du fait que nous ne savons pas si l'échantillon d'ouvrage à disposition est représentatif de l'ensemble des ouvrages du réseau situé dans l'état E_1. Il est en effet impossible de connaître la répartition des ouvrages selon les états E_1 et E_2 et de fait, il est impossible d'extraire de nos données d'interventions un échantillon représentatif.
La seconde critique concerne le choix du site pilote pour le second modèle. Outre le manque de représentativité du patrimoine, un des points faibles les plus notables est le faible taux d'instrumentation en télérelevé de la zone. Nous préconisions sur un secteur comme celui de Canéjan un taux d'instrumentation proche des 40% pour permettre une estimation fiable des pertes ; sur le Bas Cenon, même si les deux zones ne sont pas identiques, nous pouvons facilement admettre qu'un taux d'instrumentation de 10% n'est pas suffisant pour estimer précisément les pertes sur le réseau.

Chapitre 5

Conclusion et perspectives

5.1 Conclusion

L'utilisation des données de télérelevé apporte une toute nouvelle information sur le réseau. La connaissance des consommations sur un pas de temps journalier, permet de mieux gérer le réseau en réduisant les temps de détection des fuites. Nous l'avons vu dans le chapitre 3, un télérelevé généralisé, s'il est trop coûteux pour l'opérateur, n'est pas nécessaire pour connaître la consommation totale d'une population. En effet, la constitution d'un échantillon de compteurs à équiper permet de palier ce problème financier tout en garantissant une estimation fiable des consommations totales. Nous avons, tout au long du chapitre 3, développé une méthodologie d'échantillonnage adaptée au contexte opérationnel des réseaux d'eau potable. L'application d'un sondage stratifié, avec des strates définies de sorte que les groupes formés soient le plus homogènes possible, permet d'améliorer la représentativité des individus mais aussi des comportements. Ces comportements évoluant avec le temps, il est nécessaire de prendre en compte ces changements pour éviter l'obsolescence de l'échantillon. Le redressement des estimateurs, et en particulier le redressement par régression, permet de palier cet incident sans pour autant devoir investir davantage dans l'instrumentation du réseau. Ainsi, la méthode d'échantillonnage développée ne nécessite aucun investissement particulier (mise à part une analyse préliminaire pour définir les strates) mais elle est plus performante que d'autres méthodes plus "standards" (comme le sondage aléatoire simple par exemple).

L'acquisition des données de consommation à partir du télérelevé permet non seulement d'estimer les consommations d'une population sur un pas de temps journalier (ou infra-journalier) mais elle permet aussi de calculer, par différence avec le volume livré au réseau, les pertes en eau journalières (ou infra-journalières) sur le réseau. Ce volume d'eau perdu apporte à lui seul une information capitale, car il est très souvent une donnée d'entrée des indicateurs de performance, qui permettent aux opérateurs d'évaluer la dégradation de leur réseau. Cependant, il est possible d'enrichir l'information sur les pertes en décomposant la chronique des pertes selon les trois états du concept BABE : les pertes liées aux fuites invisibles et indétectables, les pertes liées aux fuites invisibles mais détectables par recherche de fuite et les pertes liées aux casses manifestes.

Le modèle de décomposition de la chronique de pertes, construit au chapitre 4, re-

pose sur deux “sous-modèles”. Le premier “sous-modèle” est un modèle semi-markovien, multi-états qui permet d’estimer, pour un ouvrage donné, la probabilité qu’il se situe dans un des états de dégradation. Ce modèle, appliqué à des données de panel, intègre des covariables expliquant la dégradation avancée de certains ouvrages. Le second “sous-modèle” est, quant à lui, un modèle de débit de fuite, permettant d’estimer pour un ouvrage, et pour un état de dégradation donné, le volume d’eau perdu. Le croisement de ces deux “sous-modèles” génère le modèle de décomposition de la chronique de pertes qui permet de calculer, à partir d’une chronique de pertes globales, le volume lié à des fuites invisibles et indétectables, le volume lié à des fuites invisibles mais détectables par recherche de fuite ainsi que le volume lié à des casses manifestes. Un tel modèle permet d’une part d’identifier la cause des pertes sur un secteur, mais appliqué à différents secteurs hydrauliques, il permettrait aussi de comparer ces secteurs et ainsi prioriser les actions à mettre en place.

5.2 Les limites et les évolutions possibles de ce travail

Les différents modèles et méthodes développés au cours de ce manuscrit ont tous une plus-value opérationnelles mais montrent aussi certaines limites. La méthodologie d’échantillonnage développée autour du cas de Canéjan, repose en partie sur le choix du taux de sondage. Or, il faut souligner que ce taux a pu être choisi uniquement parce que nous disposions des vraies données de consommation et qu’il a été possible de faire le lien entre le taux d’instrumentation et la précision de l’estimateur obtenue. La réplication directe de cette méthodologie sur un cas d’étude totalement différent de Canéjan pourrait engendrer des résultats d’estimation peu convenables (notamment en termes de précision). Une des perspectives envisageables, pour améliorer cette méthode, serait de la reproduire sur différents contrats, entièrement télérelevés et présentant des caractéristiques socio-éco-démographiques différentes (composition des ménages, des logements, répartition de l’activité économique, etc.), afin de disposer d’un panel à partir duquel il sera possible de calculer la précision des estimateurs en fonction du taux d’instrumentation. L’application sur un nouveau contrat de la méthode d’échantillonnage se fera alors à partir d’une population similaire (selon des données socio-éco-démographiques mais aussi des données d’exploitation) faisant partie du panel.

Enfin concernant la seconde partie du manuscrit et le modèle de décomposition, nous avons plusieurs fois souligné le manque d’information concernant les données d’intervention (en particulier l’état de dégradation E_2 n’est jamais observé) mais aussi la mauvaise qualité (l’imprécision) des chroniques de pertes. Une des grandes limites de ce modèle est la capacité à trouver un secteur où nous disposons de données patrimoniales et d’interventions complètes, de données de consommation (même estimées) et de données de pression en continu. Le travail réalisé sur la décomposition de la chronique de pertes pourrait être grandement améliorer en augmentant le taux d’instrumentation en télérelevé sur le secteur du Bas Cenon.

Bibliographie

Alegre, H., Hirnir, W., Baptista, J., and Parena, R. (2000). *Performance Indicators for Water Supply Services*. IWA Publishing.

Alkasseh, J., Adlan, M., I.Abustan, Aziz, H., and Hanif, A. (2013). Applying minimum night flow to estimate water loss using statistical modeling : A case study in kinta valley,malaysia. *Water Resources Management*, 27(5) :1439–1455.

Alwan, L. and Roberts, H. (1988). Time-series modeling for statistical process control. *Journal of Business & Economic Statistics*, 6(1) :87–95.

Amoatey, P., Minke, R., and Steinmetz, H. (2014). Leakage estimation in water networks based on two categories of night-time users : a case study of a developing country network. *Water Science & Technology : Water Supply*, 14(2) :329–336.

Andreou, S. (1986). *Predictive models for pipe break failures and their implications on maintenance planning strategies for deteriorating water distribution systems*. PhD thesis, MIT.

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris, France.

Babykina, G. (2011). *Modélisation statistique d'événements récurrents. Exploration empirique des estimateurs, prise en compte d'une covariable temporelle et application aux défaillances des réseaux d'eau.* PhD thesis, Ecole Doctorale de Mathématiques et Informatique.

Beaumont, J.-F., Haziza, D., and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100(3) :555–569.

Chavent, M., Kuentz, V., Labenne, A., Liquet, B., and Saracco, J. (2014). *PCAmixdata : Multivariate analysis for a mixture of quantitative and qualitative data*. R package version 2.1.

Chesneau, O. (2006). *Un outil d'aide à la maîtrise des pertes dans les réseaux d'eau potable : la modélisation dynamique de différentes composantes du débit de fuite*. PhD thesis, Université Louis Pasteur, Strasbourg I.

Clark, C. (1971). Expansive-soil effect on burried pipe. *Journal of AWWA*, 63 :424–427.

Claudio, K., Couallier, V., and Y.Le Gat (2014). Integration of time-dependent covariates in recurrent events modelling : application to failures on drinking water networks. *Journal de la Société Française de Statistiques*, 155(3) :62–77.

Cochran, W. (1977). *Sampling Techniques 3rd Edition*. Wiley Series, New York, France.

Commenges, D. and Gégout-Petit, A. (2007). Likelihood for generally coarsened observations from multi-state or counting process models. *Scandinavian Journal of Statistics*, 134(2) :432–450.

Croux, C., Gelper, S., and Mahieu, K. (2011). Robust control charts for time-series data. *Expert Systems with Applications*, 38 :13810–13815.

Dalenius, T. (1950). The problem of optimum stratification. *Skand. Akt. Tidskrift*, page 203.

Dalenius, T. and Hodges, J. (1959). Minimum variance stratification. *American Statistical Society*, 54 :88–101.

Deville, J. and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418) :376–382.

Deville, J., Särndal, C., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423) :1013–1020.

Eisenbeis, P. (1994). *Modélisation statistique de la prévision des défaillances sur les conduites d'eau potable*. PhD thesis, Université Louis Pasteur.

Fang, Y. and Zhang, J. (1999). Performance of control charts for autoregressive conditional heteroscedasticity processes. *Journal of Applied Statistics*, 26(6) :701–714.

Fanner, P. (2003). Assessing real losses, including component analysis and economic considerations : A practical approach. *Water 21 special series*.

Fantozzi, M. and Lambert, A. (2010). Legitimate night use component of minimum night flows initiative. Water Loss 2010 - Sao Paulo.

Farley, M. and Trow, S. (2003). *Losses in Water Distribution Networks - A Practitioner's Guide to Assessment, Monitoring and Control*. IWA Publishing.

Fellegi, I. (2010). Méthodes et pratiques d'enquête. `http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-fra.pdf`. Statistique Canada.

Foucher, Y., Giral, M., Soulillou, J., and Daures, J. (2010). A flexible semi-markov model for interval-censored data and goodness-of-fit testing. *Statistical Methods in Medical Research*, 19 :127–145.

Foucher, Y., M.Giral, Soulillou, J., and Daures, J. (2007). A semi-markov model for multistate and interval-censored data with multiple terminal events. application in renal transplantation. *Statistics in medecine*, 26 :5381–5393.

Gilbert, P. and Varadhan, R. (2012). *numDeriv : Accurate Numerical Derivatives*. R package version 2012.9-1.

Goulter, I. and Kanzemi, A. (1988). Spatial and temporal groupings of water main pipe breakage in winnipeg. *Canadian Journal of Civil Engineering*, 5 :91–97.

Jay, J. K., Li, J., and Valliant, R. (2007). Regroupement de cellules lors de la poststratification. *Techniques d'enquête*, 33(2) :157–170.

Kalbfleisch, J. and Lawless, J. (1985). The analysis of panel data under a markov assumption. *Journal of the american statistical association*, 80(392) :863–871.

Kalbfleisch, J. and Prentice, R. (2002). *The statistical analysis of failure time data*. New York, USA, 2nd ed. edition.

Kang, M. and Lagakos, S. (2007). Statistical methods for panel data from a semi-markov process, with application to hpv. *Biostatistics*, 8(2) :252–264.

Kleiner, Y. and Rajani, B. (2002). Forecasting variations and trends in water main breaks. *Journal of Infrastructure Systems*, 8(4) :122–131.

Kpedekpo, G. (1973). Recent advances on some aspects of stratified sample design. a review of the literature. *Metrika*, 20(1) :55–64.

Kuentz, V., Lyser, S., Candau, J., Deuffic, P., Chavent, M., and Saracco, J. (2013). Une approche par classification de variables pour la typologie d'observation : le cas d'une enquête agriculture et environnement. *Journal de la Société Française de Statistiques*, 154(2) :27–63.

Lagakos, S., Sommer, C., and Zen, M. (1978). Semi-markov model for partially censored data. *Biometrika*, 65(2) :311–317.

Lambert, A. (1994). Accounting for losses : The bursts and background concept. *Water and Environment Journal*, (2) :205–214.

Lambert, A. (2001). What do we know about pressure : leakage relationships in distribution systems ? In *IWA Conference on "System Approach to Leakage Control and Water Distribution Systems Management"*. IWA Publishing. ISBN 80-7204-197-5.

Lambert, A. (2002). International report : Water losses management and techniques. *Water Science and Technology : Water Supply*, (6) :1–20.

Lambert, A. (2009). Ten years experience in using the uarl formula to calculate infrastructure leakage index. In *Proc. of the 5th IWA Water Loss Reduction Specialist Conference*, pages 189–196.

Lambert, A., Brown, T., M.Takizawa, and Weimer, D. (1999). A review of performance indicators for real losses from water supply systems. *Aqua*, 48(6) :227–237.

Lavallée, P. and Hidiroglou, M. (1988). On the stratification of skewed population. *Techniques d'enquête*, 14 :35–45.

Le Gat, Y. (2008). Modelling the deterioration process of drainage pipelines. *Urban Water Journal*, 5(2) :97–106.

Le Gat, Y. (2009). *Une extension du processus de Yule pour la modélisation stochastique des événements récurrents*. PhD thesis, Agro Paris Tech.

Le Gat, Y. and Eisenbeis, P. (2000). Using maintenance records to forecast failures in water networks. *Urban Water*, 2 :173–181.

Lucas, J. and Saccucci, M. (1990). Exponentially weighted moving average control schemes : Properties and enhancements. *Technometrics*, 32(1) :1–12.

Melato, D., Carvalho, G., and Thornton, J. (2009). Using component analysis and infrastructure condition factor (icf) field tests to prioritize service connection replacement and reduce real losses in a sustainable manner. In *Proc. of the 5th IWA Water Loss Reduction Specialist Conference*, pages 197–205.

Montgomery, D. (2009). *Statistical Quality Control : A Modern introduction, 6th edition*. John Wiley and Sons.

Morris, R. (1967). Principal causes and remedies of water main breaks. *Journal of AWWA*, 54 :47–50.

Morrison, J. (2004). Managing leakage by district metered areas : a practical approach. *Water 21*, 6(1) :44–46.

Neyman, J. (1934). On the two aspects of representative method : the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 94(4) :558–625.

Nicolini, G. (2001). A method to define strata boundaries. Università deglei Studi di Milano.

Pilcher, R. (2003). Leak detection practices and techniques : a practical approach. *Water 21*, 5(6) :44–45.

R Development Core Team (2011). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, C. (1965). *Linear Statistical Inference and its Applications*. John Wiley and Sons, New York, USA, 2nd ed. edition.

Reynolds, M. and Chao-Wen, L. (1997). Control chart for monitoring processes with autocorrelated data. *Nonlinear Analysis, Theory, Methods & Applications*, 30(7) :4059–4067.

Rivest, L. (1999). Stratum jumpers : can we avoid them. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Roberts, S. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3) :239–250.

Røstum, J. (2000). *Statistical modelling of pipes failures in water networks.* PhD thesis, Norwegian University of Science and Technology.

Schmid, W. and Schone, A. (1997). Some properties of the ewma control chart in the presence of autocorrelation. *The Annals of Statistics*, 35(3) :1277–1283.

Scrucca, L. (2004). qcc : an r package for quality control charting and statistical process control. *R News*, 4/1 :11–17.

Serfling, R. (1968). Approximately optimal stratification. *Journal of the American Statistical Association*, 63(324) :1298–1309.

Särndal, C., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en population finie.* Dunod.

Van Zyl, J. and Clayton, C. (2007). The effects of pressure on leakage in water distribution system. In *Proc. I.C.E Water Management*, pages 104–109.

V.Barbu and N.Limnios (2008). *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications : Their Use in Reliability and DNA Analysis*. Springer Publishing Company, Incorporated, 1 edition.

Wols, B. and Van Thienen, P. (2014). Modelling the effect of climate change induced soil settling on drinking water distribution pipes. *Computers and Geotechnics*, 55 :240–247.

Woodall, W. and Faltin, W. (1993). Autocorrelated data and statistical process control. *ASQC Statistics Division Newsletter*, 13(4) :18–21.

Zhang, L. (2000). Post-stratification and calibration - a synthesis. *American Statistical Association*, 54(3) :178–184.

Annexe A

L'Entreprise Régionale Lyonnaise des Eaux Bordeaux Guyenne

L'Entreprise Régionale Bordeaux Guyenne compte actuellement 75 contrats de délégation des services d'eau potable (tout type de contrats : concession, affermage, etc.). Nous nous intéressons particulièrement ici à trois contrats en particulier :

- le contrat de concession du service d'eau potable de la Communauté Urbaine de Bordeaux (CUB),
- le contrat d'affermage du service d'eau potable de la commune de Canéjan,
- le contrat d'affermage du service d'eau potable du syndicat d'alimentation en eau de Carbon Blanc (SIAO),

Ces trois contrats sont représentées à la Figure A.1.

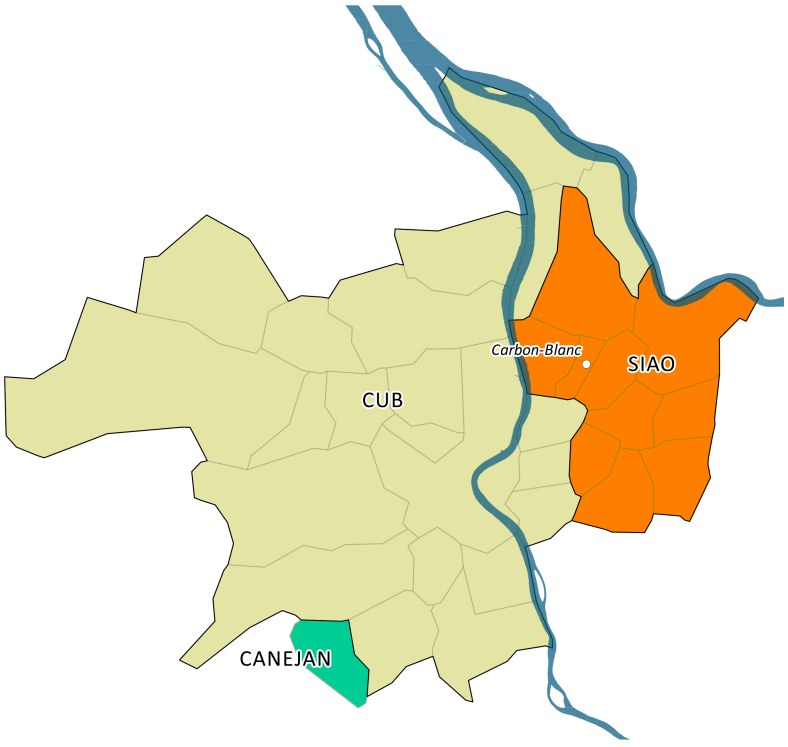


FIGURE A.1 – Implantation de Lyonnaise des Eaux dans la région bordelaise

Annexe B

Estimation de la consommation journalière en 2011

Les résultats d'estimation par sondage stratifié de la consommation journalière pour l'année 2011 ont été présentés partiellement au chapitre 3, Figure 3.7. Nous présentons à la Figure B.1 les résultats globaux qui ont été résumés à la Table 3.7.

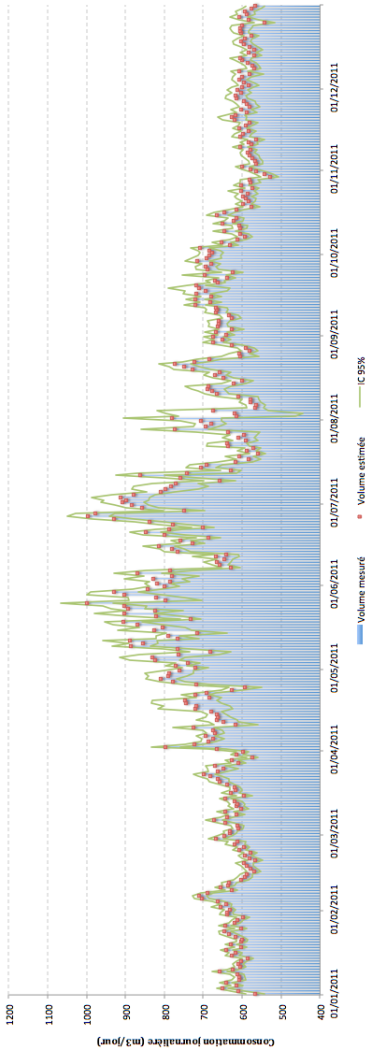


FIGURE B.1 – Estimation du total de la consommation journalière sur l'année 2011

Annexe C

Régression linéaire sur chacune des 11 strates de consommation

Comme il a été présenté en section 3.3.2 du chapitre 3, il est possible de réaliser un redressement par régression de chacun des sous-totaux au sein de chaque strate de consommation. La duplication du redressement est appropriée dans le cas où la relation entre la variable d'intérêt $Y(t)$ et la variable de redressement Z diffère d'une strate à l'autre. C'est pourquoi il paraît justifier d'analyser au préalable ces relations.

Dans la mesure où t varie dans notre cas entre le 01/01/2012 et le 01/04/2012, nous ne pouvons présenter l'intégralité des résultats, nous avons donc choisi de ne représenter que les relations entre $Y(t)$ et Z pour t = 03/02/2012. La Figure C.1 illustre d'une part la variable $Y(t)$ en fonction de Z (bleu), la régression linéaire calée sur les données de la population (vert) ainsi que celle calée sur chaque strate (rouge).

On remarque que l'ensemble des coefficients $\hat{\beta}_h$ ont une valeur proche du coefficient de régression $\hat{\beta}$ estimé sur la population. La seule exception notable est le coefficient de la strate 3. Cependant, la droite qui semble s'ajuster au mieux au nuage de points est celle calculée sur l'ensemble de la population. La droite calée sur la strate est très fortement influencée par les individus ayant une valeur importante pour Z (> 200 $m^3.an^{-1}$). L'estimation du paramètre de pente sur la population filtrée de ces individus renvoie une valeur de coefficient égale à 2.4×10^{-3}, proche du coefficient calée sur la population. De plus, le coefficient de corrélation linéaire entre $Y(t)$ et Z est de 0.45 sur la strate filtrée contre 0.38 sur la strate globale.

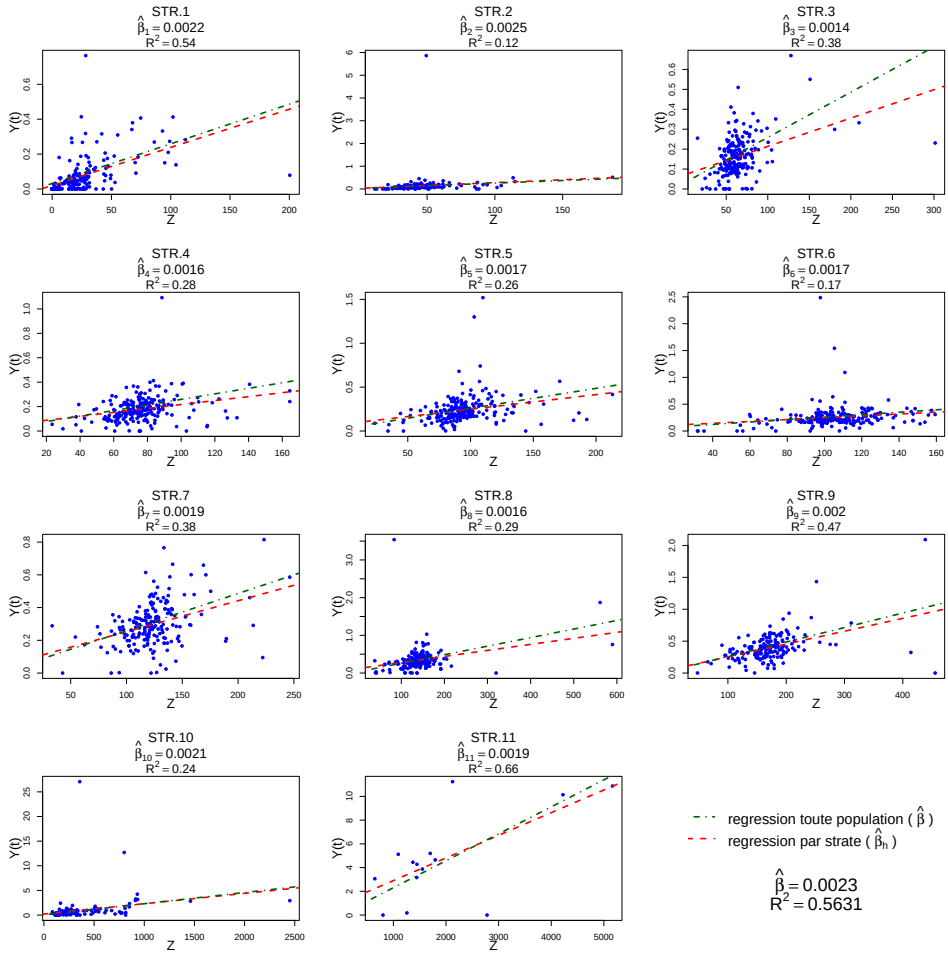


FIGURE C.1 – Régression linéaire entre la variable d'intérêt $Y(t)$ (03/02/2012) et la variable auxiliaire Z pour chaque strate (en m^3)

Annexe D

Redressement par calage

Nous avons énoncé en section 3.3.3 le fait qu'il existait pour un ensemble de distance D une fonction F associée permettant de calculer les nouveaux poids de sondage de chaque individu. Nous allons dans cette partie démontrer la relation existante entre les distances D^{α} et les fonctions F^{α} présentées au Tableau 3.14 en particulier pour la cas $\alpha = 2$. Nous avons choisi ce cas spécifique pour sa facilité de résolution. Dans ce cas on définit D ainsi :

$$D^{\alpha}(w_i, d_i) = \frac{(w_i - d_i)^2}{2d_i} \tag{D.1}$$

Le principe du redressement par calage est de recalculer des nouveaux poids w_i selon une variable auxiliaire Z. Le problème à résoudre est le suivant :

$$\begin{cases} \min \sum_s D^{\alpha}(w_i, d_i) \\ s.c. \sum_s w_i Z_i = \sum_U Z_i \end{cases} \Leftrightarrow \begin{array}{c} \min \sum_s \dfrac{(w_i - d_i)^2}{2d_i} \\ s.c. \sum_s w_i Z_i = \sum_U Z_i \end{array}$$

En utilisant la méthode d'optimisation sous contraintes de Lagrange, on définit la fonction lagrangienne par :

$$\mathscr{L}(\overrightarrow{w}, \lambda) = \sum_{i \in s} \left(\frac{(w_i - d_i)^2}{2d_i} \right) - \lambda \left(\sum_{i \in s} w_i Z_i - \sum_{i \in U} Z_i \right)$$

où $\overrightarrow{w} = (w_1, ..., w_n)^T$. La solution du problème d'optimisation s'obtient en annulant le gradient du lagrangien.

$$
\begin{aligned}
\nabla\mathscr{L}(\overrightarrow{w},\lambda)=0 \Leftrightarrow & \left\{\begin{array}{rcl} \dfrac{\partial\mathscr{L}}{\partial w_1} & = & 0 \\ & \cdots & \\ \dfrac{\partial\mathscr{L}}{\partial w_n} & = & 0 \\ \dfrac{\partial\mathscr{L}}{\partial\lambda} & = & 0 \end{array}\right. \\
\Leftrightarrow & \left\{\begin{array}{rcl} \left(\dfrac{w_1-d_1}{d_1}\right)-\lambda Z_1 & = & 0 \\ & \cdots & \\ \left(\dfrac{w_n-d_n}{d_n}\right)-\lambda Z_n & = & 0 \\ \sum_s w_i Z_i-\sum_U Z_i & = & 0 \end{array}\right. \\
\Leftrightarrow & \left\{\begin{array}{rcl} w_1 & = & d_1(1+\lambda Z_1) \\ & \cdots & \\ w_n & = & d_n(1+\lambda Z_n) \\ \sum_s w_i Z_i & = & \sum_U Z_i \end{array}\right.
\end{aligned}
\tag{D.2}
$$

On constate ainsi d'après le système d'équation (D.2) que les nouveaux poids w_i s'écrivent :

$$w_i = d_i(1+\lambda Z_i)$$

Ainsi il est possible d'associer à la distance D^α, telle que définie à l'équation (D.1), une fonction F telle que $F(t)=(1+t)$. De plus si on définit $\delta(w_i,d_i)$ telle que

$$\delta(w_i,d_i)=\frac{\partial D(w_i,d_i)}{\partial w_i}=\frac{w_i-d_i}{d_i},$$

nous avons bien

$$d_i F(.)=\delta^{-1}(.,d_i).$$

Par ailleurs, compte tenu de l'écriture de la fonction F, il est possible d'extraire une solution explicite au problème d'optimisation en continuant la résolution du système (D.2) :

$$
\begin{aligned}
\sum_s w_i Z_i=\sum_U Z_i &\Leftrightarrow \sum_s d_i(1+\lambda Z_i)Z_i=\sum_U Z_i \\
&\Leftrightarrow \sum_s d_i Z_i+\lambda\sum_s d_i Z_i^2=\sum_U Z_i \\
&\Leftrightarrow \hat{T}_{Z\pi}+\lambda\sum_s d_i Z_i^2=T_Z \\
&\Leftrightarrow \lambda\sum_s d_i Z_i^2=T_Z-\hat{T}_{Z\pi} \\
&\Leftrightarrow \lambda=\frac{T_Z-\hat{T}_{Z\pi}}{\sum_s d_i Z_i^2}
\end{aligned}
\tag{D.3}
$$

Ainsi, l'estimateur redressé par calage selon la distance D définie à l'équation D.1 s'écrit :

$$\begin{aligned}\hat{T}_{Y\,Cal^2}(t) &= \sum_s w_i Y_i(t) \\ &= \sum_s d_i(1+\lambda Z_i)Y_i(t) \\ &= \sum_s d_i Y_i(t) + \lambda \sum_s d_i Z_i Y_i(t) \\ &= \hat{T}_{Y\pi}(t) + \frac{\sum_s d_i Z_i Y_i(t)}{\sum_s d_i Z_i^2}\left(T_Z - \hat{T}_{Z\pi}\right)\end{aligned} \tag{D.4}$$

On remarque au vu de l'équation (D.4) la proximité entre les méthodes de redressement par calage et par régression.

Annexe E

Estimation de la consommation journalière 2012

Par souci de lisibilité, nous n'avons représenté que partiellement les résultats d'estimation de la consommation journalière 2012. Même si le Tableau 3.16 résume les résultats sur toute la période, nous présentons à la Figure E.1 ces mêmes résultats.

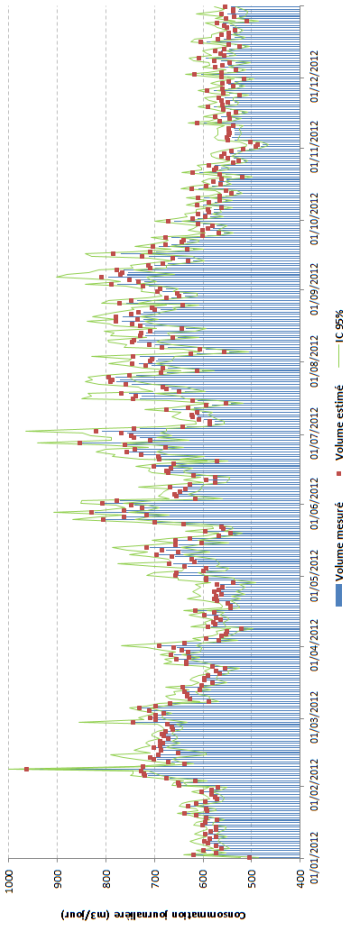


FIGURE E.1 – Estimation du total de la consommation journalière sur l'année 2012 - redressement de l'estimateur par régression

Annexe F

Le partrimoine d'eau potable de la CUB par secteur de niveau 2

TABLE F.1 – Répartition du patrimoine modélisé par secteur de niveau 2

Secteur	Linéaire modélisé	Branchement modélisés
AMELIN	74.9	1 315
BEAUREGARD/ ST-LOUIS-DE-MONTFERRAND	26.9	657
BEAUREGARD/ ST-VINCENT-DE-PAUL	31.9	377
BEAUREGARD/ZM 43 AMBES	49.6	1 052
BEGLES/ZM 60 BEGLES	139.6	10 820
BOULIAC	35.5	0
CAP ROUX/BARBUSSE	5	344
CAP ROUX/HAILLAN	0.3	0
CAP ROUX/LE HAILLAN	75.4	3 030
CAP ROUX/MERIGNAC CENTRE	104.2	4 696
CAP ROUX/MERIGNAC LA FORET	52.6	2 988
CAP ROUX/MERIGNAC SUD	91.1	3 227
CAP ROUX/MERIGNAC ZI PHARE	40.5	580
CAP ROUX/TASTA	17	847
CAP ROUX/ZM 75 TAILLAN	147.1	6 903
CAP ROUX/ZM 75 TASSIGNY	106.8	7 748
CAP ROUX/ZM 75 VERDUN	120.3	9 961
CAZEAUX	131.9	4 766
GAJAC/ST-MEDARD	112.5	5 073
GAJAC/TAILLAN	64.9	3 147
HAUT-BRION/ZM 75 POUJEAU	104.6	8 290
LINAS/BLANQUEFORT	66.6	2 324
LINAS/ZM 46 PAREMPUYRE	52.1	2 293
LINAS/ZM 75 BLANQUEFORT	52.6	1 654
PASTEUR	42.2	1 613
PAULIN BEQUET/BASSINS A FLOT	13.8	850

suite sur la prochaine page

Secteur	Linéaire modélisé	Branchement modélisés
PAULIN BEQUET/BASTIDE	65.5	3 862
PAULIN BEQUET/BELCIER	18.5	793
PAULIN BEQUET/CENTRE	287.5	24 014
RIVE DROITE/BAS CENON	16.6	1 227
RIVE DROITE/BAS LORMONT	11.9	640
RIVE DROITE/HAUT CENON	57.5	2 369
RIVE DROITE/HAUT FLOIRAC	49.6	2 335
RIVE DROITE/HAUT LORMONT	48.2	1 756
RIVE DROITE/YVRAC	12.3	768
ROUQUET/PESSAC CENTRE	88.4	4 046
ROUQUET/PESSAC MAGONTY	128.5	5 740
ROUQUET/PESSAC SAIGE	98.7	2 960
ROUQUET/ROUQUET 75	2.2	0
SAINT-AUBIN	158.1	6 014
SAUSSETTE/COUHINS	11.8	612
SAUSSETTE/VILLENAVE SUD	83.6	3 840
SAUSSETTE/ZM 60 VILLENAVE	64.6	2 342
SAUSSETTE/ZM 75 THOUARS	98.4	4 658

Annexe G

Écriture du gradient pour la log vraisemblance du modèle GompitZ

$$\frac{\partial \ln L(\theta)}{\partial \theta_i} = \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{\frac{\partial S_{y_{ij}}(t)}{\partial \theta_i} - \frac{\partial S_{y_{ij}-1}(t)}{\partial \theta_i}}{S_{y_{ij}}(t) - S_{y_{ij}-1}(t)} = \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{\frac{\partial S_{y_{ij}}(t)}{\partial \theta_i} - \frac{\partial S_{y_{ij}-1}(t)}{\partial \theta_i}}{\mathbb{P}(Y(t) = y_{ij})}$$

$$\nabla S_k(\theta) \begin{cases} \frac{\partial S_k(t)}{\partial \alpha_k} = & -\exp(-\exp(\alpha_k + Z_0\beta_0 + t\exp(Z_1\beta_1))) \\ & \exp(\alpha_k + Z_0\beta_0 + t\exp(Z_1\beta_1)) \\ \frac{\partial S_k(t)}{\partial \alpha_l} = & 0 \qquad \text{si} \qquad k \neq l \\ \frac{\partial S_k(t)}{\partial \beta_0} = & -Z_0\exp(-\exp(\alpha_k + Z_0\beta_0 + t\exp(Z_1\beta_1))) \\ & \exp(\alpha_k + Z_0\beta_0 + t\exp(Z_1\beta_1)) \\ \frac{\partial S_k(t)}{\partial \beta_1} = & -tZ_1\exp(Z_1\beta_1)\exp(-\exp(\alpha_k + Z_0\beta_0 + t\exp(Z_1\beta_1))) \\ & \exp(\alpha_k + Z_0\beta_0 + t\exp(Z_1\beta_1)) \end{cases}$$

Cas spécifique où $\theta_i = \alpha_k$:

	$\frac{\partial S_{y_{ij}}(t)}{\partial \alpha_1} - \frac{\partial S_{y_{ij}-1}(t)}{\partial \alpha_1}$	$\frac{\partial S_{y_{ij}}(t)}{\partial \alpha_2} - \frac{\partial S_{y_{ij}-1}(t)}{\partial \alpha_2}$
$y_{ij} = 1$	$\frac{\partial S_1(t)}{\partial \alpha_1}$	0
$y_{ij} = 2$	$-\frac{\partial S_1(t)}{\partial \alpha_1}$	$\frac{\partial S_2(t)}{\partial \alpha_2}$
$y_{ij} = 3$	0	$-\frac{\partial S_2(t)}{\partial \alpha_2}$

Annexe H

Résultats d'estimation des paramètres des modèles de dégradation

Les différents modèles développés au chapitre 4 ont été appliqués à notre jeu de données. Les tableaux H.1 à H.4 résument pour chacun des matériaux les résultats d'estimation des paramètres utilisés dans les modèles.

TABLE H.1 – Paramètres estimés par matériau pour le modèle GompitZ

Facteur	FD/AC	FG/AM	PVC/PE	PE Bleu	PE Noir
α_1	-5.864	-4.300	-5.060	-3.716	-3.138
α_2	-5.959	-4.350	-5.127	-3.872	-3.186
Diamètre	-0.003	-0.004	-0.004		
Collier				0.423	
Fourreau				-1.674	
Lotissement					0.072
Intercept	-1.084	-1.506	-0.482	-9.603	0.256
Longueur	0.491	0.502	0.491		
Pression (minimale)	0.113				
Pression (variation)		0.321			
Pression (modulation)					-35.518
Matériau (Génération)		-0.863		9.874	0.420
Matériau (Indicatrice)			0.211		
Climat (Froid)	0.017	0.045			
Climat (Chaud)				0.026	0.039

TABLE H.2 – Paramètres estimés par matériau pour le modèle LEYP

Facteur	FD/AC	FG/AM	PVC/PE	PE Bleu	PE Noir
α	8.198	1.714	7.540	4.188	2.788
δ	0.964		1.240	1.348	2.481
Intercept	-4.653	-3.249	-3.605	-0.150	0.781
Longueur	0.635	0.604	0.564		
Diamètre	-0.003	-0.003			

suite sur la prochaine page

Facteur	FD/AC	FG/AM	PVC/PE	PE Bleu	PE Noir
Pression (minimale)	0.065	0.039			
Pression (modulation)				-0.128	-0.266
Collier				0.274	
Sol Argileux				0.231	0.201
Matériau (Génération)		-0.107			
Matériau (Indicatrice)			0.400		
Climat (Froid)	0.104	0.171			
Climat (Chaud)				0.060	0.085

TABLE H.3 – Paramètres estimés par matériau pour le modèle de risques compétitifs

Facteur	FD/AC	FG/AM	PVC/PE	PE Bleu	PE Noir
λ_2	0.007	0.378	0.011	0.085	1.265
λ_3	0.181	0.452	0.115	1.169	1.799
δ_2	1.432	6.499	1.455	1.104	6.267
δ_3	2.264	4.410	1.698	1.664	4.292
Longueur	0.734	0.876	0.736		
Diamètre	-0.002	-0.003			
Collier				0.556	
Sol Argileux				0.196	0.193
Pression (minimale)	0.049	0.167			
Matériau (Génération)		-2.781			
Matériau (Indicatrice)			0.296		
Nb casses passées	1.480	0.231	0.989	0.873	0.536
Climat (Chaud)				0.067	0.088

TABLE H.4 – Paramètres estimés par matériau pour le modèle multi-états de dégradation

Facteur	FD/AC	FG/AM	PVC/PE	PE Bleu	PE Noir
δ_{12}	1.222	3.903	1.816	5.016	1.345
$\delta_{22'}$	1.253	0.821	0.185	0.271	0.383
δ_{23}	1.041	0.646	0.420	0.775	0.805
η_{12}	11.429	9.031	0.580	0.450	6.636
$\eta_{22'}$	0.011	717.221	4.749	2.012	12.465
η_{23}	0.001	91.223	0.015	0.021	0.001
Intercept	-2.475	5.978			-1.272
Diamètre	-0.003	-0.001			
Longueur	0.480	0.580			0.662
Matériau (Génération)		-2.355			
Nb casses passées		0.234			0.353
Collier			1.580		
Sol Argileux				0.176	