An extension of the Yule process for the stochastic modelling of recurrent events. Application to the failures of pressure water mains.

Yves Le Gat

N° /__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

Docteur

de

l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité : Sciences de l'eau (Option Statistique)

présentée et soutenue publiquement
par

Yves Le Gat

le 11 décembre 2009

Une extension du processus de Yule pour la modélisation stochastique des événements récurrents

Application aux défaillances de canalisations d'eau sous pression

Directeur de thèse : ***Daniel Commenges***
Codirecteur de thèse : ***Éric Parent***
Travail réalisé au Cemagref de Bordeaux

Devant le jury :

Jean-Jacques Boreux	**Chef de Travaux**	***Université de Liège***	**Président**
Alain Mailhot	**Professeur**	***INRS***	**Rapporteur**
Jérôme Saracco	**Professeur**	***Université de Bordeaux***	**Rapporteur**
Bernard Brémond	**Directeur de Recherche**	***Cemagref***	**Examinateur**
Daniel Commenges	**Directeur de Recherche**	***INSERM***	**Examinateur**
Éric Parent	**Ingénieur en Chef PEF**	***AgroParisTech***	**Examinateur**
Jean-Philippe Torterotot	**Chef de Département**	***Cemagref***	**Examinateur**

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech) est un Grand Etablissement dépendant du Ministère de l'Agriculture et de la Pêche, composé de l'INA PG, de l'ENGREF et de l'ENSIA (décret n° 2006-1592 du 13 décembre 2006)

Résumé

Après une présentation du processus de Yule dans sa forme classique, une extension en est proposée à but de modélisation stochastique des événements récurrents. Dans le cadre de ce processus étendu, une formule analytique est démontrée par récurrence pour la distribution conditionnelle du nombre d'événements susceptibles de se produire dans un intervalle de temps donné, sachant le nombre d'événements subis avant le début de l'intervalle ; la connaissance de cette distribution est importante pour modéliser un processus réel observé uniquement sur un intervalle de temps, qui est borné et ne débute pas nécessairement à $t = 0$. Il est montré que le cas particulier d'une dépendance linéaire entre l'intensité du processus et le rang de l'événement conduit à la distribution binomiale négative du processus de comptage. Dans ce cas linéaire, conduisant au *Linear Extended Yule Process* (LEYP), la fonction de vraisemblance des paramètres du modèle connaissant une séquence d'événements observés est construite, qui permet d'estimer ces paramètres. Est ensuite établie la forme particulière (binomiale négative) que prend la probabilité conditionnelle du nombre d'événements susceptibles de se produire dans un intervalle de temps donné, connaissant le nombre d'événements qui se sont produits dans un intervalle de temps antérieur ; ce résultat est indispensable pour valider le modèle, et effectuer des prévisions. La fonction de vraisemblance n'est cependant valide qu'à la condition que le système étudié soit toujours observable. Dans le cas d'un système à durée de vie limitée aléatoirement par le nombre d'événements subis, il est montré comment les données d'observation peuvent être biaisées par le phénomène de la survie sélective, et comment prendre en compte la survie du système dans la construction de la fonction de vraisemblance.
Faisant suite à ces investigations théoriques, l'estimation pratique des paramètres est étudiée par des simulations sur ordinateur. Il est montré à cet effet comment générer une séquence aléatoire d'événement distribués selon un LEYP, comment estimer les paramètres en maximisant leur fonction de vraisemblance, et enfin comment effectuer des prévisions de nombres d'événements, et effectuer la validation de ces prévisions en établissant une courbe de performance prédictive du modèle. Le modèle LEYP est enfin mis en oeuvre sur des exemples réels de données de défaillance, et son efficacité pratique est mise en évidence.
Mots clés : Processus de Yule ; Extension linéaire du processus de Yule ; Evénements récurrents ; Fenêtre d'observation ; Distribution binomiale négative ; Biais de survie sélective ; Performance prédictive d'un modèle.

Abstract

After the Yule Process in its classical form has been presented, an extension is proposed that aims at stochastically modelling recurrent events. In the frame of this extended process, the analytical form of the conditional distribution of the possible number of events that may occur in a given time interval given the number of events that already occurred before the beginning of the interval is proven by induction; knowing this distribution is important when seeking to model an actual process only observed within a bounded time interval that not necessarily starts at $t = 0$. The particular case of a process intensity that linearly depends on the rank of the event is shown to generate a counting process whose distribution is negative binomial. In this linear case leading to the *Linear Extended Yule Process* (LEYP), the likelihood function of the model parameters given a sequence of observed events is built, and allows to estimate these parameters. The number of events that may occur in a time interval given the number of events observed in a previous time interval is then proven to also have a negative binomial distribution; this result is essential to validate the model and perform predictions. The likelihood function is nevertheless valid provided the system under study can always be observed. In the case of a system whose life time is randomly limited according to the number of past events, it is shown how the data may be biased by the selective survival phenomenon, and how to account for the system survival when building the likelihood function.

After these theoretical investigations, the practical estimation of the model parameters is studied with random computer simulations. It is shown to that end how to generate a random sample of events distributed according to a LEYP, how to estimate the parameters by maximizing their likelihood function, and finally how to perform predictions of the number of events, and how to cross-validate these predictions by building a predictive performance curve. The LEYP model is lastly used to analyse real failure datasets, that demonstrate its practical efficiency.

Key words: Yule Process ; Linear Extension of the Yule Process ; Recurrent events ; Observation window ; Negative binomial distribution ; Selective survival bias ; Model predictive performance.

Avant-propos

Ce travail de recherche, mené au sein de l'équipe NetWater du Groupement de Bordeaux du Cemagref, est pleinement redevable aux échanges et collaborations avec mes collègues, et particulièrement avec :

- Bernard Brémond, Directeur de Recherche, pionnier de la recherche française en gestion patrimoniale des réseaux et mentor de notre équipe, qui m'a accueilli en 1995 au Cemagref, et m'a soutenu et guidé jusqu'au bout de ce cursus doctoral,
- Patrick Eisenbeis, Ingénieur en Chef, auteur lui-même en 1994 d'un travail de thèse précurseur sur ce même sujet des défaillances de canalisations,
- Eddy Renaud, Ingénieur en Chef, qui a mené le développement du logiciel Casses, grâce auquel le modèle LEYP peut être mis à disposition des praticiens de la gestion patrimoniale des réseaux d'eau, portant ainsi ce processus de recherche jusqu'à son stade ultime du transfert,

ainsi qu'à l'enseignement de biostatistique, puis aux nombreux conseils avisés prodigués par Daniel Commenges, Directeur de Recherche INSERM à l'Université de Bordeaux 2, en matière de processus markoviens, processus de comptage et construction de leur vraisemblance.

Les échanges avec Éric Parent, Directeur du Laboratoire MORSE de l'institut AgroParisTech, qui m'a accueilli au sein de l'école doctorale Abiès, ainsi qu'avec Jean-Jacques Boreux, Chef de travaux à l'université de Liège, qui m'a honnoré de sa participation à mon comité de thèse ainsi qu'à mon jury, ont fort utilement alimenté mes réflexions.

L'initiation des recherches sur le modèle LEYP doit beaucoup à l'implication internationale de notre équipe. La visite en 1998 de Jon Røstum, alors doctorant à l'Université de Trondheim, et les fructueuses discussions concernant l'application du modèle NHPP aux défaillances de canalisations, ont fortement influencé mon travail. L'implication en 2000 de Bernard Brémond dans le jury de thèse de Geneviève Pelletier, alors Doctorante à l'Université Laval de Québec sous la direction d'Alain Mailhot, Professeur à l'INRS, a porté à mon attention les travaux de son équipe qui ont profondément marqué mes réflexions sur la modélisation du processus de défaillance.

La participation de notre équipe entre 2002 et 2005 au projet européen Care-W a été l'occasion de confronter les points de vue entre chercheurs issus de laboratoires ayant des cultures scientifiques différentes, ainsi qu'avec des ingénieurs impliqués dans la gestion patrimoniale de réseaux d'eau d'importantes collectivités ; particulièrement, les échanges avec Jean-Philippe Torterotot, Chef de Département au Cemagref, et qui m'a en outre fait l'honneur de participer à mon jury de thèse, Sveinung Saegrov, Professeur à l'Université de Trondheim et coordonnateur du projet Care-W, Aïtor Ibarrola et Sébastien Apothéloz, Ingénieurs à EauService-Lausanne, Pascal Le Gauffre, Professeur à l'INSA de Lyon, Caty Werey, Ingénieur au Laboratoire GSP du Cemagref à Strasbourg, Matthew Poulton, Ingénieur du Bureau d'Étude WTSim de Bordeaux,

Raimund Herz, Professeur à l'Université de Dresde, Rolf Baur et Ingo Kropp, Ingénieurs du Bureau d'Étude Baur+Kropp de Dresde, Annie Vanrenterghem-Raven, Professeur au Polytechnic Institute de New York, ont aiguillonné mes réflexions concernant le versant appliqué de cette recherche.

Les discussions suivies, avec Michel Guillon, Cyril Leclerc, Khaled Odeh, Ingénieurs à Suez-Environnement, avec Daniel Revol, François Pinson, Sébastien Buttoudin, Ingénieurs à Veolia-Eau, sur la mise en oeuvre de la modélisations des défaillances ont considérablement alimenté mes réflexions, y compris au plan théorique.

Ma collaboration avec Genia Babykina, Doctorante au sein de notre UR qui travaille sur la prise en compte de facteurs de risque de défaillance dépendants du temps, et avec Vincent Couallier, Maître de Conférence à l'Université de Bordeaux 2 et Jérôme Saracco, Professeur à l'Université de Bordeaux, qui dirigent son travail, m'ont permis à la fois d'approfondir ma façon de voir, et de prendre du recul par rapport au sujet de mes recherches.

Les commentaires sur mon manuscrit de thèse, ainsi que les critiques et suggestions formulés par les Professeurs Alain Mailhot et Jérôme Saracco, qui m'ont fait l'honneur d'accepter d'être rapporteurs de mon travail, ont grandement contribué à la qualité de la version finale de ce document.

Que tous trouvent ici l'expression sincère de ma gratitude.

Bordeaux, janvier 2010

« *Mais, ne proposant cet écrit que comme une histoire, ou, si vous l'aimez mieux que comme une fable en laquelle, parmi quelques exemples qu'on peut imiter, on en trouvera peut-être aussi plusieurs autres qu'on aura raison de ne pas suivre, j'espère qu'il sera utile à quelques uns, sans être nuisible à personne, et que tous me sauront gré de ma franchise.* »

René Descartes, *Discours de la Méthode* (1637)

« *Et détournant mes yeux de ce vide avenir*
En moi-même je vois tout le passé grandir »

Guillaume Apollinaire, *Le cortège* (Alcools, 1913)

Table des matières

Table des figures

Liste des tableaux

Présentation générale

Les exemples de systèmes subissant des événements récurrents abondent, tant dans le domaine de la Santé Publique, que dans celui de la fiabilité des systèmes techniques. Parmi ces exemples pratiques, nombreux sont ceux qui exhibent une double tendance à voir la fréquence de répétition des événements augmenter avec l'âge du patient ou de l'objet technique, ainsi qu'avec le nombre d'événements déjà subis. Cette augmentation avec l'âge est bien modélisée par le classique Processus de Poisson Non Homogène (NHPP), particulièrement étudié par [Lawless, 1987], et dont des exemples intéressants de mise en oeuvre sont décrits par [Samset, 1994] dans le cas de pompes de centrales électriques, et par [Røstum, 2000] dans le cas de canalisations d'eau potable. La prise en compte de l'augmentation avec le nombre d'événements passés, mise en avant par [Eisenbeis, 1994] dans le contexte des canalisations d'eau, justifie en revanche un investissement théorique auquel le présent travail souhaiterait contribuer, tout en gardant en vue le souci d'appliquer les résultats à la modélisation du risque de défaillance des canalisations d'eau potable.

Le présent ouvrage se divise donc en une première partie théorique présentant une tentative de synthèse du modèle proposé par [Eisenbeis, 1994] et du NHPP, sous la forme d'une extension linéaire du Processus de Yule, et en une seconde partie appliquée, consacrée à la mise en oeuvre de ce nouvel outil pour modéliser le risque de défaillance sur les conduites composant un réseau d'eau potable.

Première partie

Théorie

Chapitre 1

Introduction

La théorie des Processus Stochastiques est le cadre mathématique « naturel » d'étude de la récurrence d'événements aléatoires de même nature. Comme présenté dans [Cook et Lawless, 2002], cette étude peut être abordée selon deux angles d'attaque théoriquement équivalents, et consistant respectivement à modéliser :

– soit la distribution du temps écoulé entre deux événements successifs,
– soit la distribution du nombre d'événements advenant dans un intervalle de temps.

La voie choisie par [Eisenbeis, 1994] relève de la première catégorie. La présentation « classique » de [Ross, 1983] relève de la seconde. L'extension linéaire du Processus de Yule non homogène (nommée *LEYP* dans la suite) vise à fonder un modèle de risque de défaillances qui cumule les avantages du NHPP et du modèle Eisenbeis. Cela implique une mise au point théorique, centrée sur la notion de Processus de comptage, qui fait l'objet de cette première partie.

La présentation de cette première partie ayant une portée générale, l'entité concernée par les défaillances répétées sera dénommée « objet technique » ; ce terme sera cependant remplacé par « canalisation » ou « conduite » lorsque le contexte fera plus spécifiquement référence aux défaillances affectant un réseau d'eau.

1.1 Notation

L'ensemble du document obéit aux conventions de notation suivantes :

– $\mathbb{N}$ et $\mathbb{N}^*$ sont respectivement l'ensemble des entiers naturels $\{0, 1, 2, \ldots, \infty\}$ et l'ensemble des entiers naturels strictement positifs $\{1, 2, \ldots, \infty\}$;
– $\mathbb{R}$, $\mathbb{R}_+$ et $\mathbb{R}_+^*$ sont les intervalles réels $]-\infty, +\infty[$, $[0, +\infty[$ et $]0, +\infty[$;
– $\Pr\{A\}$ et $\Pr\{A \mid B\}$ dénotent respectivement la probabilité de l'événement A, et la probabilité conditionnelle de A sachant B ;
– $\Pr\{A \cap B\}$ et $\Pr\{A, B\}$ notent tous deux la probabilité de l'occurrence simultanée des événements A et B ;
– $t \in \mathbb{R}_+$ est une variable temporelle positive représentant en pratique l'âge de l'objet technique ;
– $N(t) \in \mathbb{N}$ est la fonction en escalier (à valeurs entières) de comptage des défaillances ;
– $\mathrm{d}N(t)$ est la différentielle de $N(t)$, *i.e.* $\mathrm{d}N(t) = 1$ si $N(t)$ connaît un saut à t, $\mathrm{d}N(t) = 0$

sinon ;
- $\mathscr{N}_{[a,t[}$ représente la σ-algèbre auto-excitatrice générée par le processus $N(t)$ dans $[a, t[$;
- $\mathscr{N}_{t-}$ dénote $\mathscr{N}_{[0,t[}$;
- $\boldsymbol{Z}$ est le vecteur des facteurs de risque propres à l'objet technique (aussi nommés covariables) ;
- $\mathscr{F}_{[a,t[} = \mathscr{N}_{[a,t[} \vee \sigma(\boldsymbol{Z})$ dénote l'information sur le processus $\mathscr{N}_{[a,t[}$ augmentée de celle sur les covariables $\boldsymbol{Z}$, ou plus techniquement la plus petite σ-algèbre qui contient tous les événements composés à partir des éléments des σ-algèbres $\mathscr{N}_{[a,t[}$ et $\sigma(\boldsymbol{Z})$;
- $\lambda(t)$ est une fonction réelle positive bornée sur un compact, et son intégrale est $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$;
- $\mathrm{E}X$ et $\mathrm{E}(X \mid A)$ notent respectivement l'espérance de la variable aléatoire X et son espérance conditionnelle sachant A ;
- $\mathrm{Var}(X)$ représente la variance de la v.a. X ;
- $\mathcal{U}_E$ dénote la distribution uniforme sur l'ensemble E ;
- $\mathcal{U}_{[0,1]}$ dénote en particulier la distribution uniforme dans l'intervalle $[0, 1]$;
- $\mathcal{N}(\mu, \sigma^2)$ dénote la distribution normale d'espérance μ et de variance σ^2 ;
- $\mathcal{P}o(\mu)$ dénote la distribution de Poisson d'espérance $\mu \in \mathbb{R}_+$;
- $\mathcal{NB}(k, p)$ dénote la distribution binomiale négative de paramètres $k \in \mathbb{R}_+^*$ et $p \in [0, 1]$;
- $\chi^2(k)$ dénote la distribution de Chi2 à $k \in \mathbb{N}^*$ degrés de liberté ;
- $L(\boldsymbol{\theta})$ représente la vraisemblance du processus théorique de paramètre $\boldsymbol{\theta}$ connaissant une séquence de défaillances observées ;
- $\prod$ note l'opérateur de produit intégral, qui joue pour les produits un rôle analogue à celui que l'opérateur $\int$ joue pour les sommes ;
- la fonction indicatrice $\mathrm{I}(p)$ de la proposition p prend la valeur 1 si p est vraie, 0 sinon ;
- l'opérateur min renvoie le minimum d'une collection de valeurs ;
- l'opérateur max renvoie le maximum d'une collection de valeurs.

L'ensemble des lignes de calcul constituant la preuve d'une proposition sera close par le symbole □ justifié à droite ; de même les lignes de texte exprimant une remarque seront en caractères italiques et closes par le symbole △ justifié à droite.

1.2 Cadre théorique général

Nous adoptons dans l'ensemble de ce document l'approche théorique développée par P.K. Andersen, Ø. Borgan, R.D. Gill et N. Keiding dans leur indispensable ouvrage de référence *Statistical Models Based on Counting Processes* [Andersen *et al.*, 1993]. Considérons un objet technique avec des enregistrements de maintenance disponibles dans l'intervalle de temps $[a, b]$, avec m défaillances observées aux instants $0 \leqslant a \leqslant t_1 < \ldots < t_j < \ldots < t_m \leqslant b < +\infty$. Par convention, nous considérons que $t_0 = a$ et $t_{m+1} = b$. L'origine $t = 0$ de l'axe du temps est fixée à la date de mise en service de l'objet technique. La séquence de défaillances est supposée être une réalisation des variables aléatoires $T_j : j = 1, \ldots, m$ (par convention : $T_0 = 0$). Nous définissons le processus de comptage $N(t)$ comme la fonction entière, continue à droite et limitée à gauche,

qui s'incrémente d'une unité à chaque T_j :

$$\begin{aligned}&\forall t \in \{T_j : j = 1, \ldots, m\}, \quad \mathrm{d}N(t) = 1\\ &\forall t \in]T_j, T_{j+1}[: j = 0, \ldots, m, \quad \mathrm{d}N(t) = 0\end{aligned}$$

L'intensité de $N(t)$, *i.e.* la densité de probabilité d'un saut à t, est conditionnelle à la σ-algèbre auto-excitatrice tronquée à gauche $\mathscr{N}_{[a,t[}$:

$$\mathscr{N}_{[a,t[} = \sigma\left(N(s) - N(a)\right)_{s\in[a,t[}$$

$\mathscr{N}_{[a,t[}$ peut être vue comme la connaissance que nous avons du processus depuis le début de son observation jusque juste avant t. De façon heuristique, l'intensité de $N(t)$ peut être définie comme :

$$\Pr\left\{\mathrm{d}N(t) = 1 \mid \mathscr{N}_{[a,t[}\right\} = \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)$$

Nous supposons qu'au plus une défaillance peut survenir à un instant donné, et que le processus de comptage ne peut pas « exploser », *i.e.* reste fini à tout instant fini :

$$\begin{aligned}&\forall t \in \mathbb{R}_+,\\ &\begin{cases} \Pr\left\{\mathrm{d}N(t) > 1 \mid \mathscr{N}_{[a,t[}\right\} = 0\\ \Pr\left\{N(t) < \infty \mid \mathscr{N}_{[a,t[}\right\} = 1\end{cases}\end{aligned}$$

1.3 Les modèles Eisenbeis et NHPP

Dans le modèle de [Eisenbeis, 1994], dénommé par la suite « modèle Eisenbeis », les délais séparant les événements successifs sont des variables aléatoires X_j définies sur $\mathbb{R}_+$, indexées par le rang d'occurrence $j \in \mathbb{N}$ des événements, et distribuées selon des lois de Weibull dont les paramètres de position μ_j et de forme δ_j sont fonctions de j. La fonction de répartition de X_j s'écrit ainsi :

$$\forall x \in \mathbb{R}_+, \forall j \in \mathbb{N}, \Pr\left\{X_j \leq x \mid \mu_j, \delta_j\right\} = 1 - \exp\left(-x^{\delta_j} e^{\mu_j}\right)$$

Le paramètre de position est en outre défini comme une combinaison linéaire de variables explicatives, de nature qualitative ou numérique continue, caractérisant l'objet technique ou son environnement de fonctionnement (dans le contexte technique du modèle Eisenbeis, il s'agit de caractéristiques des canalisations d'eau telles que diamètre, longueur, position sous chaussée ou trottoir, nature du sol encaissant, *etc.*), et dont les valeurs composent le vecteur de covariables $\boldsymbol{Z}$: $\mu_j = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_j$, où $\boldsymbol{\beta}_j$ est un vecteur de paramètres. Le modèle appartient donc à la classe des modèles dits *à risques proportionnels*, ou encore *modèles de Cox*, car proposés initialement par [Cox, 1972]. Les composantes des vecteurs $\boldsymbol{Z}$ et $\boldsymbol{\beta}_j$ sont par convention indicées de 0 à q, où q est le nombre effectif de covariables (une variable numérique compte pour une covariable effective, une variable qualitative à m modalités pour $m - 1$ covariables effectives). En fixant $Z_0 = 1$, $e^{\beta_{j0}}$ représente alors le *risque de base*.

Il est cependant possible de reformuler un tel modèle en termes de *processus de comptage* $N(t)$ du nombre d'événements subis par le système pendant l'intervalle $[0, t]$:

Définition 1.1. *Le modèle Eisenbeis est défini par le système d'équations :*

$$\begin{array}{l} \forall t \in]T_{j-1}, T_j], \forall j \in \mathbb{N}^*, \\ \left\{ \begin{array}{l} N(0) = 0 \\ \mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) = j-1\right) = \delta_j (t - T_{N(t-)})^{\delta_j - 1} e^{\boldsymbol{Z}^T \boldsymbol{\beta}_j} \mathrm{d}t \end{array} \right. \end{array}$$

avec par convention $T_0 = 0$ à la pose de la canalisation.

Afin de ne pas avoir à estimer un nombre déraisonnable de paramètres, [Eisenbeis, 1994] propose de simplifier la dépendance de δ_j et $\boldsymbol{\beta}_j$ sur j en regroupant les valeurs de j en 3 strates :

- Strate I pour $j \in \{1\}$,
- Strate II pour $j \in \{2, 3, 4\}$
- et Strate III pour $j \in \{5, 6, \ldots\}$,

et en fixant de plus $\delta_{III} = 1$ dans la 3ème strate.

En comparaison le NHPP, tel qu'exposé par [Ross, 1983], peut se définir par :

Définition 1.2. *Le Processus de Poisson Non Homogène est défini par le système d'équations :*

$$\begin{array}{l} \forall t \in \mathbb{R}_+, \\ \left\{ \begin{array}{l} N(0) = 0 \\ \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right) = \mathrm{E}\left(\mathrm{d}N(t)\right) = \lambda(t)\mathrm{d}t \end{array} \right. \end{array}$$

Un cas particulier, présenté par [Lawless, 1987] comme facilement utilisable en pratique, est obtenu en posant l'intensité : $\lambda(t) = \delta t^{\delta-1} e^{\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{\beta}}$.

Les définitions respectives 1.1 et 1.2 des modèles Eisenbeis et NHPP font ressortir une différence essentielle : dans le modèle Eisenbeis l'intensité du processus dépend fortement de *l'ordre de la défaillance*, alors que l'intensité du NHPP est gouvernée par *l'âge du processus*.

1.4 Autres approches pour les défaillances de canalisations d'eau

Dans le domaine d'application des défaillances répétées affectant les canalisations d'eau de nombreuses approches sont rapportées dans la littérature. Un excellent panorama concernant les publications depuis 1979 en a été dressé par [Kleiner et Rajani, 2001]. Il est à noter qu'à l'exception des travaux centrés sur les délais inter-défaillances, le cadre théorique des processus stochastiques, et plus spécifiquement celui des processus de comptage, n'est jamais utilisé. Cette tendance semble vouloir perdurer, puisque les travaux les plus récents tels que ceux de [Debón *et al.*, 2010] et [Yamijala *et al.*, 2009] mettent en avant l'intérêt des modèles linéaires généralisés appliqués à la probabilité d'occurrence d'au moins une défaillance dans un intervalle de temps de quelques années et en utilisant la distribution de Poisson dans le cas de [Debón *et al.*, 2010], ou dans un intervalle de quelques mois et en utilisant la loi logistique dans le cas de [Yamijala *et al.*, 2009].

1.5 Pourquoi le Processus de Yule ?

La définition 1.1 du modèle Eisenbeis met en évidence une limitation importante : l'estimation des paramètres à l'aide d'un jeu de données observées n'est possible que si les objets techniques sont observés depuis leur mise en service ; dans le cas contraire où l'observation est restreinte à un intervalle de temps $[a, b]$ avec $a > 0$, le rang des événements est inconnu et le modèle est inapplicable. Des applications pratiques ont cependant été conduites en consentant « l'approximation » consistant à prendre $t = 0$ en début de fenêtre d'observation, et en introduisant le logarithme de l'ordre de la défaillance ainsi que l'âge à la défaillance précédente en covariables ; les résultats, somme toute intéressants, sont rapportés par [Le Gat, 1999] et [Le Gat et Eisenbeis, 2000], et montrent un avantage certain sur le modèle NHPP, quant à la détection des objets techniques les plus à risque. Le modèle Eisenbeis permet en effet de bien les détecter, et ainsi de prioriser les réhabilitations. Les calculs de prévision du nombre de défaillances par objet technique ont cependant toujours montré une tendance gênante à la sous- ou surestimation, et ce couramment d'un facteur 2 ; en comparaison, le NHPP détecte mal les objets techniques les plus à risque, mais produit des prévisions non biaisées en moyenne. La mise en oeuvre numérique des calculs de prévision requiert en outre une technique lourde en temps de calcul de simulation de Monte Carlo, destinée à pallier l'impossibilité analytique de convoluer des distributions de Weibull ; le NHPP permet au contraire des calculs de prévisions très simples à implémenter et peu coûteux en temps de calcul.

La présente investigation du Processus de Yule est donc largement motivée par le besoin de disposer d'un modèle alliant le bon pouvoir de détection des objets les plus à risque du modèle Eisenbeis, à la simplicité de mise en oeuvre et l'absence de biais dans les prédictions du NHPP. Le processus recherché doit impliquer une augmentation de la fréquence d'événements avec l'âge ainsi qu'avec le nombre d'événements passés. L'idée d'enchaîner des délais inter-événementiels exponentiellement distribués, dont le paramètre varie avec l'ordre de l'événement est présentée dans [Le Gauffre *et al.*, 2001], qui fait référence à la distribution de Furry (parfois aussi dénommée de Yule-Furry). La possibilité de traiter des défaillances dont l'observation est limitée au sein d'intervalles de temps $[a, b]$ ne débutant pas nécessairement en $a = 0$, en explicitant les probabilités $\Pr\{N(b) - N(a) = m \mid N(a) = j\}$, puis en prenant l'espérance sur j est exposée dans [Pelletier, 1999] (bien que les notions de vraisemblances conditionnelle et marginale n'y soient pas explicites). L'idée d'un traitement utilisant le Processus de Yule est donc largement redevable à ces deux dernières références. La base théorique du Processus de Yule est bien décrite par [Ross, 1983].

1.6 Organisation de la première partie

Après un rappel au chapitre 2 concernant la loi binomiale négative et le processus de Yule, est présentée au chapitre 3 la forme plus générale du processus de naissance non homogène (NHBP), en insistant particulièrement en section 3.3 sur la forme que prend la probabilité conditionnelle du nombre d'événements dans un intervalle temporel, connaissant le nombre d'événements subis entre $t = 0$ et le début de l'intervalle. Ce résultat est ensuite appliqué au chapitre 4 au cas particulier où l'intensité du NHBP prend une forme analytique linéaire en le nombre d'événements passés, ce qui définit le processus que nous dénommons « LEYP » (pour

Linear Extension of the Yule Process) dans la suite ; l'expression analytique de la probabilité binomiale négative d'observer un nombre donné d'événements dans un intervalle de temps donné, connaissant le nombre d'événements observés dans un intervalle antérieur est établie dans des configurations de plus en plus générales en section 4.1, ainsi que le pendant marginal de ces probabilités en sous-sections 4.1.2 et 4.1.5. Le chapitre 5 établit la forme analytique de la vraisemblance d'un processus théorique LEYP connaissant les séquences de défaillances subies par un échantillon aléatoire de canalisations de caractéristiques connues observées dans des intervalles donnés ; la forme analytique des dérivées premières et secondes de la log-vraisemblance par rapport aux paramètres du modèle LEYP est elle aussi établie, permettant l'implémentation de l'algorithme d'estimation de Nelder-Mead. Le chapitre 6 présente une extension du modèle LEYP, destinée à prendre en compte un possible biais de survie sélective dans les données servant au calage du modèle, par la considération d'un processus bidimensionnel, impliquant deux types concurrents de défaillances, celles conduisant à une réparation, et celles suivies de la mise hors service de l'objet technique (censure à droite de l'observation, dépendante du processus de défaillances).

Chapitre 2

Le Processus de Yule et la loi Binomiale Négative

Sont brièvement présentés dans ce chapitre le processus de Yule et la loi binomiale négative ; bien que les résultats qui suivent soient classiques en calcul des probabilités, nous avons cependant jugé utile de les exposer en détail. Notre but est là de familiariser le lecteur avec les manipulations analytiques relatives aux probabilités binomiales négatives, qui seront utiles pour comprendre les démonstrations des prochains chapitres sans avoir préalablement à se replonger dans des cours de calcul des probabilités.

Le processus de Yule a été initialement conçu comme un modèle simple de croissance dans le temps d'une population ne comptant initialement qu'un seul individu, au sein de laquelle les individus ont un taux de mortalité nulle, et un taux de reproduction $\lambda > 0$ constant dans le temps. Dans la littérature anglophone, ce processus est ainsi souvent dénommé *Pure Birth Process*. Il est commode, suivant [Ross, 1983], de considérer le processus de Yule comme le processus de comptage $N(t)$ du nombre d'individus composant la population au temps t , selon la :

Définition 2.1. *Le processus de Yule d'intensité $\lambda \in \mathbb{R}_+$ est défini par le système d'équations :*

$$\forall t \in \mathbb{R}_+, \forall j \in \mathbb{N}^*,$$
$$\begin{cases} N(0) = 1 \\ \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right) = \mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) = j\right) = j\lambda \mathrm{d}t \end{cases}$$

Le processus $N(t)$ possède les deux propriétés importantes suivantes :
- $N(t)$ est un processus markovien, la probabilité d'observer un événement entre t et $t + \mathrm{d}t$ ne dépendant que du nombre d'événements déjà observés juste avant t ;
- au plus un événement peut être observé dans un intervalle de temps infinitésimal.

Une conséquence remarquable est que la distribution de $N(t)$ sur $\mathbb{N}^*$ est géométrique :

$$\forall t \in \mathbb{R}_+, \forall j \in \mathbb{N} \ : \Pr\{N(t) = j + 1\} = e^{-\lambda t}(1 - e^{-\lambda t})^j \tag{2.1}$$

Si l'effectif initial de la population est $k \geqslant 1$, la distribution de $N(t)$ sur $\mathbb{N} \cap [k, +\infty[$ s'obtient en sommant les nombres aléatoires de descendants des k fondateurs, c'est à dire en convoluant k fois la distribution géométrique précédente ; il en résulte une distribution binomiale négative

$\mathcal{NB}(k, e^{-\lambda t})$:

$$\forall t \in \mathbb{R}_+, \forall j \in \mathbb{N} \ : \ \Pr\{N(t) = j + k\} = \binom{j+k-1}{k-1} e^{-k\lambda t}(1 - e^{-\lambda t})^j \tag{2.2}$$

Il est intéressant de rapprocher ce mode de génération d'une loi binomiale négative de celui classiquement présenté dans la littérature du calcul des probabilités (*cf.* par exemple [Renyi, 1966]). Soit l'événement $A(j, k)$ défini par $j + k$ expériences aléatoires de Bernouilli, identiques et indépendantes, de probabilité de succès p, au cours desquelles sont observés j échecs et k succès, dont le dernier lors de la dernière épreuve. Soit la variable aléatoire discrète J prenant la valeur $j \in \mathbb{N}$ lorsque l'événement $A(j, k)$ se réalise. La fonction de probabilité de J est alors :

$$\Pr\{J = j\} = \binom{j+k-1}{k-1} p^k (1 - p)^j \tag{2.3}$$

qui est simplement la probabilité conjointe de k succès et j échecs multipliée par le nombre de façons de placer les $k - 1$ premiers succès parmi les $j + k - 1$ épreuves qui précèdent le $k^{\text{ème}}$ succès final. J suit donc bien une distribution binomiale négative : $J \sim \mathcal{NB}(k, p)$.

La définition de la loi binomiale négative $\mathcal{NB}(k, p)$, où k est un entier, s'étend sans difficulté à celle de la loi $\mathcal{NB}(\theta, p)$, où θ est un réel positif. La fonction de probabilité de la variable aléatoire entière $J \in \mathbb{N} \sim \mathcal{NB}(\theta, p)$ prend alors la forme :

$$\Pr\{J = j\} = \frac{\Gamma(\theta + j)}{\Gamma(\theta) j!} p^\theta (1 - p)^j \tag{2.4}$$

Remarquons que la forme (2.4) est cohérente avec celle de (2.3) du fait de la propriété de la fonction Gamma :

$$\forall n \in \mathbb{N} \ : \ \Gamma(n + 1) = n!$$

Mentionnons enfin les formules permettant de calculer l'espérance et la variance de la variable aléatoire entière $J \sim \mathcal{NB}(\theta, p)$. Pour les établir nous utiliserons la propriété suivante de la fonction $\Gamma(.)$:

$$\forall z \in \mathbb{R}_+^*, \ \Gamma(z + 1) = z\Gamma(z) \tag{2.5}$$

Proposition 2.1. *L'espérance de la v.a. $J \sim \mathcal{NB}(\theta, p)$ est :*

$$\mathrm{E}(J) = \frac{\theta(1 - p)}{p}$$

Preuve :

Nous utilisons (2.5) et le changement de variable $l = j - 1$:

$$\begin{aligned} \mathrm{E}(J) &= \sum_{j=0}^{\infty} j \Pr\{J = j\} = \sum_{j=0}^{\infty} j \frac{\Gamma(\theta + j)}{\Gamma(\theta) j!} p^\theta (1 - p)^j \\ &= \sum_{j=1}^{\infty} (\theta + j - 1) \frac{\Gamma(\theta + j - 1)}{\Gamma(\theta)(j - 1)!} p^\theta (1 - p)(1 - p)^{j-1} \end{aligned}$$

$$= \theta(1-p)\sum_{l=0}^{\infty}\frac{\Gamma(\theta+l)}{\Gamma(\theta)l!}p^{\theta}(1-p)^{l} + (1-p)\sum_{l=0}^{\infty} l\frac{\Gamma(\theta+l)}{\Gamma(\theta)l!}p^{\theta}(1-p)^{l}$$
$$= \theta(1-p) + (1-p)\mathrm{E}(J)$$

D'où :

$$\mathrm{E}(J) - (1-p)\mathrm{E}(J) = \theta(1-p)$$

Et donc :

$$\mathrm{E}(J) = \frac{\theta(1-p)}{p}$$

□

Proposition 2.2. *La variance de la v.a. $J \sim \mathcal{NB}(\theta, p)$ est :*

$$\mathrm{Var}(J) = \frac{\theta(1-p)}{p^2}$$

Preuve :

$$\mathrm{Var}(J) = \sum_{j=0}^{\infty} j^2 \Pr\{J = j\} - \mathrm{E}(J)^2$$

Nous utilisons (2.5) et le changement de variable $l = j - 1$:

$$\mathrm{Var}(J) + \mathrm{E}(J)^2 = \sum_{j=0}^{\infty} j^2 \frac{\Gamma(\theta+j)}{\Gamma(\theta)j!}p^{\theta}(1-p)^{j}$$
$$= \sum_{j=1}^{\infty} j\frac{\Gamma(\theta+j)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j}$$
$$= \sum_{j=1}^{\infty}(j-1)\frac{\Gamma(\theta+j)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j} + \sum_{j=1}^{\infty}\frac{\Gamma(\theta+j)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j}$$
$$= \sum_{j=1}^{\infty}(j-1)(\theta+j-1)\frac{\Gamma(\theta+j-1)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j} + \sum_{j=1}^{\infty}(\theta+j-1)\frac{\Gamma(\theta+j-1)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j}$$
$$= \sum_{j=1}^{\infty}\theta(j-1)\frac{\Gamma(\theta+j-1)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j} + \sum_{j=1}^{\infty}(j-1)^2\frac{\Gamma(\theta+j-1)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j}$$
$$+ \sum_{j=1}^{\infty}\theta\frac{\Gamma(\theta+j-1)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j} + \sum_{j=1}^{\infty}(j-1)\frac{\Gamma(\theta+j-1)}{\Gamma(\theta)(j-1)!}p^{\theta}(1-p)^{j}$$
$$= \sum_{l=0}^{\infty}\theta(1-p)l\frac{\Gamma(\theta+l)}{\Gamma(\theta)l!}p^{\theta}(1-p)^{l} + \sum_{l=0}^{\infty}(1-p)l^2\frac{\Gamma(\theta+l)}{\Gamma(\theta)l!}p^{\theta}(1-p)^{l}$$

$$+\sum_{l=0}^{\infty}\theta(1-p)\frac{\Gamma(\theta+l)}{\Gamma(\theta)l!}p^{\theta}(1-p)^{l}+\sum_{l=0}^{\infty}(1-p)l\frac{\Gamma(\theta+l)}{\Gamma(\theta)l!}p^{\theta}(1-p)^{l}$$
$$=\theta(1-p)\mathrm{E}(J)+(1-p)\left(\mathrm{Var}(J)+\mathrm{E}(J)^{2}\right)+\theta(1-p)+(1-p)\mathrm{E}(J)$$

D'où :

$$\mathrm{Var}(J)-(1-p)\mathrm{Var}(J)=\theta(1-p)\mathrm{E}(J)+(1-p)\mathrm{E}(J)^{2}+\theta(1-p)+(1-p)\mathrm{E}(J)-\mathrm{E}(J)^{2}$$

Ainsi, en utilisant la proposition 2.1 :

$$p\mathrm{Var}(J)=p\mathrm{E}(J)^{2}+(1-p)\mathrm{E}(J)^{2}+p\mathrm{E}(J)+(1-p)\mathrm{E}(J)-\mathrm{E}(J)^{2}$$

Et donc :

$$\mathrm{Var}(J)=\mathrm{E}(J)/p$$

□

Chapitre 3

Processus de naissance non homogène

Dans le but de proposer une extension du processus de Yule adaptée à la modélisation des événements récurrents, le système d'équations de la définition 2.1 est modifié à trois niveaux. L'événement d'intérêt étant de façon très générale une défaillance, il est premièrement plus conforme à la réalité de considérer qu'aucun événement n'a encore été observé jusqu'à la mise en service du système à $t = 0$, et donc de prendre $N(0) = 0$. La définition 2.1 du processus est ensuite étendue en autorisant l'intensité du processus à varier avec le temps : $\lambda = \lambda(t)$. Cette extension a été proposée par [Chang *et al.*, 2002], dans le cadre du processus de Yule non homogène (« *Time-Dependent Yule Process* »).

Il est cependant intéressant, dans le cadre de notre étude de considérer dans un premier temps une relation entre l'intensité du processus et le rang j de l'événement plus générale que la proportionnalité directe, qui définit le processus communément dénommé *de pure naissance* (*Pure Birth Process*) ou *de naissance simple* (*Simple Birth Process*) dans la littérature (*cf.*par exemple [Bharucha-Reid, 1997] ou [Sen et Balakrishnan, 1999]). Le processus considéré dans ce chapitre, ayant aussi une intensité dépendante du temps, sera dénommé *processus de naissance non homogène*, abrégé en *NHBP* (pour *Non Homogeneous Birth Process*) dans la suite.

Définition 3.1. *Le NHBP est défini par le système d'équations :*

$$\begin{array}{l} \forall t \in \mathbb{R}_+, \forall j \in \mathbb{N} \ : \\ \left\{ \begin{array}{l} N(0) = 0 \\ \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right) = \mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) = j\right) = a(j)\lambda(t)\mathrm{d}t \end{array} \right. \end{array}$$

Dans un premier temps, la fonction $a(.)$ sera prise comme une fonction réelle strictement positive quelconque de j :

$$a(j) = \alpha_j \quad \text{avec :} \quad \alpha_j \in \mathbb{R}_+^*, \quad \forall j,k \in \mathbb{N} : \quad j \neq k \Rightarrow \alpha_j \neq \alpha_k, \quad \text{et} \quad \alpha_0 = 1 \tag{3.1}$$

A l'instar de l'intensité portée précédemment à la définition 2.1, l'intensité de la définition 3.1, de par le terme $a\left(N(t-)\right)$ confère au processus $N(t)$ la propriété d'être markovien.

D'après le théorème de décomposition de Doob-Meyer (*cf.*[Andersen *et al.*, 1993]) le processus $N(t)$ peut s'écrire comme la somme d'un processus prédictible appelé compensateur (le « modèle ») $A(t)$ et d'une martingale $M(t)$ (« bruit »imprédictible d'espérance nulle) :

$$N(t) = A(t) + M(t)$$

où :

$$A(t) = \int_0^t a\left(N(u-)\right)\lambda(u)\mathrm{d}u$$

3.1 Définition générale de l'intensité

La densité de probabilité d'un saut d'une unité du processus de comptage à l'instant t dépend de la σ-algèbre $\mathscr{N}_{t-}$ des trajectoires possibles de la fonction de comptage $N(t)$ entre 0 et $t-$, l'instant juste avant t. La famille croissante continue à droite des σ-algèbres $(\mathscr{N}_s, s < t)$ est appelée filtration, ou souvent de façon plus parlante « historique » ; elle représente tout ce qu'un observateur sait du processus de $t = 0$ jusqu'à $t-$. Le processus $N(t)$ est dit « naturellement adapté » à cette filtration. Suivant la présentation faite par [Andersen *et al.*, 1993], l'intensité du processus est l'espérance conditionnelle sachant $\mathscr{N}_{t-}$ de la différentielle de $N(t)$: $\mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right)$.

Avec la définition 3.1, l'intensité ne dépend que de la valeur prise par le processus juste avant t :

$$\mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right) = \mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) = j\right) = \alpha_j \lambda(t)\mathrm{d}t$$

3.2 L'intensité de Yule-Weibull-Cox

Dans les applications pratiques, plusieurs systèmes sont observés simultanément, chacun étant caractérisé par les valeurs des q covariables rassemblées dans le vecteur $\boldsymbol{Z}$. Il s'avère alors intéressant en pratique d'écrire la fonction d'intensité du processus comme le produit de trois facteurs :

- le facteur de Yule, assurant que l'intensité dépend du nombre d'événements passés,
- le facteur de Weibull, assurant que l'intensité est une fonction puissance (*power-law* dans la littérature anglophone) de l'âge du processus,
- le facteur de Cox, assurant que l'intensité est modulée de façon proportionnelle par les covariables.

Cela justifie la dénomination d'intensité de « Yule-Weibull-Cox », qui fait l'objet de la :

Définition 3.2. *L'intensité de Yule-Weibull-Cox est définie par :*

$$\begin{aligned}
& \mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) = j\right) = \alpha_j \delta t^{\delta-1} e^{\boldsymbol{Z}^T\boldsymbol{\beta}}\mathrm{d}t \\
\textit{avec :}\quad & \alpha_j \in \mathbb{R}_+^*, \quad \forall j,k \in \mathbb{N} : \quad j \neq k \Rightarrow \alpha_j \neq \alpha_k, \quad \alpha_0 = 1 \\
& \delta \in [1, +\infty[\\
& \boldsymbol{Z}^T = \begin{pmatrix} 1 & Z_1 & Z_2 & \dots & Z_q \end{pmatrix} \\
& \boldsymbol{\beta}^T = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \dots & \beta_q \end{pmatrix}
\end{aligned}$$

La présence d'un 1 en première composante de $\boldsymbol{Z}$ et du β_0 correspondant dans $\boldsymbol{\beta}$ spécifie l'intensité de base e^{β_0} qui permet de poser la condition $\alpha_0 = 1$ dans (3.1) sans restriction de

généralité. Le modèle ainsi obtenu pour plusieurs systèmes est donc un modèle de régression, qui permet d'inférer après estimation des paramètres, en appliquant un test statistique adéquat (*cf.* section 5.2), si l'effet des covariables sur l'intensité du processus est significatif ou non.

La valeur de β_j s'interprète comme l'effet de la covariable Z_j sur l'intensité de casse. Dans le cas où Z_j est une covariable indicatrice d'une propriété du système considéré, *i.e.* $Z_j = 1$ si le système possède la propriété, $Z_j = 0$ sinon ; nous rencontrerons fréquemment ce type de covariable qualitative dans les applications aux conduites d'eau en seconde partie de ce document, par exemple la position de la conduite sous chaussée ou sous trottoir. Si nous considérons deux conduites qui ne diffèrent que par la valeur de Z_j, et posons $\mu = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta} - \beta_j Z_j$. Le rapport des intensités, noté *RR*, des deux conduites, souvent dénommé « risque relatif » s'écrit :

$$RR = \frac{\Pr\left\{\mathrm{d}N(t) = 1 \mid N(t-) = j, Z_j = 1\right\}}{\Pr\left\{\mathrm{d}N(t) = 1 \mid N(t-) = j, Z_j = 0\right\}} = \frac{\alpha_j \delta t^{\delta-1} e^{\mu+\beta_j Z_j} \mathrm{d}t}{\alpha_j \delta t^{\delta-1} e^{\mu} \mathrm{d}t} = e^{\beta_j}$$

Le coefficient de régression β_j est donc ici simplement le logarithme naturel du risque relatif. Si de plus, comme nous le verrons à la section 5.2, β_j peut être estimé par $\hat{\beta}_j$, avec une précision mesurée par l'écart-type d'estimation $\sigma_{\hat{\beta}_j}$, et que l'estimateur utilisé est asymptotiquement normalement distribué, alors $\exp(\hat{\beta}_j \pm q_\alpha \sigma_{\hat{\beta}_j})$ donne les bornes d'un intervalle de confiance de $\hat{\beta}_j$ de probabilité α (la valeur la plus fréquemment utilisée en pratique est $\alpha = 0.95$, correspondant à $q_\alpha \simeq 1.96$).

Si Z_j est une covariable quantitative continue (par exemple le diamètre de la conduite), et que les deux conduites ne diffèrent que par $Z_j = z_1$ et $Z_j = z_2$, alors :

$$RR = e^{\beta_j (z_1 - z_2)}$$

Si enfin Z_j est le logarithme naturel d'une variable quantitative continue positive, telle que la longueur de la conduite par exemple, et que les deux conduites ne diffèrent que par $Z_j = \ln l_1$ et $Z_j = \ln l_2$, alors :

$$RR = \left(\frac{l_1}{l_2}\right)^{\beta_j}$$

3.3 Distribution conditionnelle du processus de comptage

Dans cette section est envisagé le cas général du NHBP d'intensité conforme à la définition 3.1 et à l'équation (3.1) par :

Définition 3.3. *L'intensité du NHBP est notée :*

$$\mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) = j\right) = \alpha_j \lambda(t) \mathrm{d}t$$

avec : $\forall t \in \mathbb{R}_+, \lambda(t) \in \mathbb{R}_+$

Nous utiliserons en outre dans la suite la notation :

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u \tag{3.2}$$

Nous allons établir une formule explicite pour calculer la probabilité conditionnelle du nombre d'événements susceptibles de se produire dans l'intervalle de temps $[t, t+s]$, où $t, s \in \mathbb{R}_+$, connaissant le nombre d'événements déjà subis jusqu'au début de l'intervalle. Suivant la méthode exposée par [Ross, 1983] dans le cas du NHPP, nous fixons t et j, et introduisons la notation :

$$Q_m(s) = \Pr\{N(t+s) - N(t) = m \mid N(t-) = j\} \tag{3.3}$$

Montrons tout d'abord que $Q_m(s)$ est solution d'un équation différentielle ordinaire linéaire. C'est l'objet de la :

Proposition 3.1. *La probabilité conditionnelle $Q_m(s)$ est solution de l'équation différentielle ordinaire linéaire du premier ordre :*

$$\begin{aligned}&\forall m \in \mathbb{N}^*,\ \forall s \in \mathbb{R}_+ :\\ &\mathrm{d}Q_m(s)/\mathrm{d}s + \alpha_{j+m}\lambda(t+s)Q_m(s) = \alpha_{j+m-1}\lambda(t+s)Q_{m-1}(s)\end{aligned} \tag{3.4}$$

avec la condition initiale :

$$Q_0(s) = \exp\left(-\alpha_j[\Lambda(t+s) - \Lambda(t)]\right) \tag{3.5}$$

Preuve :

Commençons par établir la validité de la condition initiale (3.5).

Pour $m = 0$, et t et j étant fixés, posons :

$$\begin{aligned}&Q_0(s + \mathrm{d}s)\\ &= \Pr\{N(t+s+\mathrm{d}s) - N(t) = 0 \mid N(t) = j\}\\ &= \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 0, N(t+s) - N(t) = 0 \mid N(t) = j\}\end{aligned}$$

Comme $\Pr\{N(t) = j\} \neq 0$ et $\Pr\{N(t+s) - N(t) = 0, N(t) = j\} \neq 0$, nous avons :

$$\begin{aligned}&Q_0(s + \mathrm{d}s)\\ &= \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 0 \mid N(t+s) - N(t) = 0, N(t) = j\} \times\\ &\quad \times \Pr\{N(t+s) - N(t) = 0 \mid N(t) = j\}\end{aligned}$$

$N(t)$ étant un processus markovien :

$$\Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 0 \mid N(t+s) - N(t) = 0, N(t) = j\}$$

$$= \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 0 \mid N(t+s) = j\}$$

D'où :

$$Q_0(s+\mathrm{d}s) = \Big(1 - \alpha_j \lambda(t+s)\mathrm{d}s\Big)\, Q_0(s)$$

Et donc :

$$\begin{aligned} &Q_0(s+\mathrm{d}s) - Q_0(s) = -\alpha_j \lambda(t+s) Q_0(s)\mathrm{d}s \\ &\Longrightarrow \quad \mathrm{d}Q_0(s)/Q_0(s) = -\alpha_j \lambda(t+s)\mathrm{d}s \quad \Longrightarrow \quad \mathrm{d}\ln Q_0(s) = -\alpha_j \lambda(t+s)\mathrm{d}s \end{aligned}$$

Puis en intégrant :

$$\int_0^s \mathrm{d}\ln Q_0(u) = -\alpha_j \int_0^s \lambda(t+u)\mathrm{d}u \qquad \text{avec :} \quad Q_0(0) = 1$$

Nous obtenons finalement :

$$Q_0(s) = \exp\Big(-\alpha_j[\Lambda(t+s) - \Lambda(t)]\Big)$$

Démontrons maintenant la validité de la proposition (3.4).

Posons pour tout $m \geq 1$, t et j étant fixés :

$$\begin{aligned} &Q_m(s+\mathrm{d}s) = \Pr\{N(t+s+\mathrm{d}s) - N(t) = m \mid N(t) = j\} \\ &= \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 0, N(t+s) - N(t) = m \mid N(t) = j\} \\ &\quad + \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 1, N(t+s) - N(t) = m-1 \mid N(t) = j\} \\ &= \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 0 \mid N(t+s) - N(t) = m, N(t) = j\} \times \\ &\quad \times \Pr\{N(t+s) - N(t) = m \mid N(t) = j\} \\ &\quad + \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 1 \mid N(t+s) - N(t) = m-1, N(t) = j\} \times \\ &\quad \times \Pr\{N(t+s) - N(t) = m-1 \mid N(t) = j\} \end{aligned}$$

$N(t)$ étant un processus markovien :

$$\begin{aligned} &\Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 0 \mid N(t+s) - N(t) = m, N(t) = j\} \\ &= \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 0 \mid N(t+s) = j+m\} \end{aligned}$$

et :

$$\begin{aligned} &\Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 1 \mid N(t+s) - N(t) = m-1, N(t) = j\} \\ &= \Pr\{N(t+s+\mathrm{d}s) - N(t+s) = 1 \mid N(t+s) = j+m-1\} \end{aligned}$$

D'où :

$$Q_m(s+\mathrm{d}s) = \Big(1 - \alpha_{j+m}\lambda(t+s)\mathrm{d}s\Big)\, Q_m(s) + \Big(\alpha_{j+m-1}\lambda(t+s)\mathrm{d}s\Big)\, Q_{m-1}(s)$$

Et donc :

$$\mathrm{d}Q_m(s)/\mathrm{d}s + \alpha_{j+m}\lambda(t+s)Q_m(s) = \alpha_{j+m-1}\lambda(t+s)Q_{m-1}(s) \qquad \square$$

Remarque 3.1. *Pour établir la validité de* (3.5)*, nous avons utilisé la propriété :*

$$\begin{aligned} &A, B, C \text{ étant des événements tels que } \Pr\{C\} \neq 0 \text{ et } \Pr\{B \cap C\} \neq 0, \\ &\Pr\{A \cap B \mid C\} = \Pr\{A \mid B \cap C\} \times \Pr\{B \mid C\} \end{aligned}$$

qui s'obtient aisément en considérant :

$$\Pr\{A \cap B \mid C\} = \frac{\Pr\{A \cap B \cap C\}}{\Pr\{C\}}$$

et

$$\Pr\{A \mid B \cap C\} \times \Pr\{B \mid C\} = \frac{\Pr\{A \cap B \cap C\}}{\Pr\{B \cap C\}} \times \frac{\Pr\{B \cap C\}}{\Pr\{C\}} \qquad \triangle$$

La proposition 3.1 est un cas particulier des équations différentielles de Kolmogorov exposées par exemple par [Bharucha-Reid, 1960] dans le cadre du traitement des Processus markoviens à valeurs discrètes en temps continu.

La forme analytique de la solution générale de l'équation différentielle ordinaire linéaire (3.4) est suggérée par la convolution de distributions exponentielles dont les paramètres diffèrent deux à deux, exposée dans [Cox, 1962].

Cela nous conduit à la :

Proposition 3.2. *La probabilité conditionnelle qu'un système, obéissant au processus de naissance non homogène d'intensité donnée par la définition 3.3, subisse m événements dans l'intervalle* $[t, t+s]$ *sachant que* $N(t) = j$ *est :*

$$\forall m \in \mathbb{N},\ \forall s \in \mathbb{R}_+ : \quad Q_m(s) = \left(\prod_{k=0}^{m-1} \alpha_{j+k}\right) \sum_{k=0}^{m} \frac{e^{-\alpha_{j+k}[\Lambda(t+s)-\Lambda(t)]}}{\prod_{l=0, l\neq k}^{m} (\alpha_{j+l} - \alpha_{j+k})} \tag{3.6}$$

où $\alpha_0 = 1$ *et* $j \neq k \Rightarrow \alpha_j \neq \alpha_k$.

Preuve :

La solution générale de l'équation différentielle ordinaire linéaire du premier ordre (3.4) s'obtient en posant :

$$v(s) = e^{-\int_0^s \alpha_{j+m}\lambda(t+u)\mathrm{d}u} \quad \text{et} \quad w(s) = \int_0^s \frac{\alpha_{j+m-1}\lambda(t+u)Q_{m-1}(u)}{v(u)}\,\mathrm{d}u$$

D'où la solution :

$$\begin{aligned}
&Q_m(s) = v(s)w(s) \\
&= e^{\alpha_{j+m}[\Lambda(t)-\Lambda(t+s)]} \int_0^s \alpha_{j+m-1}\lambda(t+u)Q_{m-1}(u)e^{\alpha_{j+m}[\Lambda(t+u)-\Lambda(t)]}\mathrm{d}u \\
&= e^{-\alpha_{j+m}\Lambda(t+s)} \int_0^s \alpha_{j+m-1}\lambda(t+u)Q_{m-1}(u)e^{\alpha_{j+m}\Lambda(t+u)}\mathrm{d}u
\end{aligned}$$

Montrons que l'équation de récurrence (3.6) est vraie pour $m = 1$:

$$\begin{aligned}
&Q_1(s) = e^{-\alpha_{j+1}\Lambda(t+s)} \int_0^s \alpha_j\lambda(t+u)Q_0(u)e^{\alpha_{j+1}\Lambda(t+u)}\mathrm{d}u \\
&= e^{-\alpha_{j+1}\Lambda(t+s)} \int_0^s \alpha_j\lambda(t+u)e^{\alpha_{j+1}\Lambda(t+u)-\alpha_j[\Lambda(t+u)-\Lambda(t)]}\mathrm{d}u \\
&= e^{-\alpha_{j+1}\Lambda(t+s)+\alpha_j\Lambda(t)} \frac{\alpha_j}{\alpha_{j+1}-\alpha_j} \int_0^s \mathrm{d}[e^{(\alpha_{j+1}-\alpha_j)\Lambda(t+u)}] \\
&= \alpha_j\left(\frac{e^{\alpha_j[\Lambda(t)-\Lambda(t+s)]}}{\alpha_{j+1}-\alpha_j} + \frac{e^{\alpha_{j+1}[\Lambda(t)-\Lambda(t+s)]}}{\alpha_j-\alpha_{j+1}}\right)
\end{aligned}$$

qui est bien (3.6) pour $m = 1$.

Supposons que (3.6) est vraie à l'ordre $m-1$. On a alors :

$$\begin{aligned}
Q_m(s) &= e^{-\alpha_{j+m}\Lambda(t+s)} \int_0^s \alpha_{j+m-1}\lambda(t+u)\left(\prod_{k=0}^{m-2}\alpha_{j+k}\right)\times \\
&\qquad \times \sum_{k=0}^{m-1} \frac{e^{\alpha_{j+k}[\Lambda(t)-\Lambda(t+u)]}}{\prod_{l=0,l\neq k}^{m-1}(\alpha_{j+l}-\alpha_{j+k})} e^{\alpha_{j+m}\Lambda(t+u)}\mathrm{d}u \\
&= \left(\prod_{k=0}^{m-1}\alpha_{j+k}\right)e^{-\alpha_{j+m}\Lambda(t+s)}\times \\
&\quad \times \sum_{k=0}^{m-1} \frac{e^{\alpha_{j+k}\Lambda(t)}}{\prod_{l=0,l\neq k}^{m-1}(\alpha_{j+l}-\alpha_{j+k})} \int_0^s \lambda(t+u)e^{(\alpha_{j+m}-\alpha_{j+k})\Lambda(t+u)}\mathrm{d}u \\
&= \left(\prod_{k=0}^{m-1}\alpha_{j+k}\right)e^{-\alpha_{j+m}\Lambda(t+s)}\times \\
&\quad \times \sum_{k=0}^{m-1} \frac{e^{\alpha_{j+k}\Lambda(t)}}{\prod_{l=0,l\neq k}^{m-1}(\alpha_{j+l}-\alpha_{j+k})} \int_0^s \frac{\mathrm{d}[e^{(\alpha_{j+m}-\alpha_{j+k})\Lambda(t+u)}]}{\alpha_{j+m}-\alpha_{j+k}}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\prod_{k=0}^{m-1} \alpha_{j+k}\right) e^{-\alpha_{j+m}\Lambda(t+s)} \times \\
&\quad \times \sum_{k=0}^{m-1} \frac{e^{\alpha_{j+k}\Lambda(t)} \left(e^{(\alpha_{j+m}-\alpha_{j+k})\Lambda(t+s)} - e^{(\alpha_{j+m}-\alpha_{j+k})\Lambda(t)}\right)}{\prod_{l=0,l\neq k}^{m} (\alpha_{j+l} - \alpha_{j+k})} \\
&= \left(\prod_{k=0}^{m-1} \alpha_{j+k}\right) \sum_{k=0}^{m-1} \frac{e^{\alpha_{j+k}[\Lambda(t)-\Lambda(t+s)]} - e^{\alpha_{j+m}[\Lambda(t)-\Lambda(t+s)]}}{\prod_{l=0,l\neq k}^{m} (\alpha_{j+l} - \alpha_{j+k})} \\
&= \left(\prod_{k=0}^{m-1} \alpha_{j+k}\right) \times \\
&\quad \times \left(\sum_{k=0}^{m-1} \frac{e^{\alpha_{j+k}[\Lambda(t)-\Lambda(t+s)]}}{\prod_{l=0,l\neq k}^{m} (\alpha_{j+l} - \alpha_{j+k})} - e^{\alpha_{j+m}[\Lambda(t)-\Lambda(t+s)]} \sum_{k=0}^{m-1} \frac{1}{\prod_{l=0,l\neq k}^{m} (\alpha_{j+l} - \alpha_{j+k})} \right)
\end{aligned}
$$

Utilisons l'identité (A.1) démontrée en Appendice A :

$$
\sum_{k=0}^{m} \frac{1}{\prod_{l=0,l\neq k}^{m} (\alpha_l - \alpha_k)} = 0
$$

Il vient finalement :

$$
\begin{aligned}
Q_m(s) &= \left(\prod_{k=0}^{m-1} \alpha_{j+k}\right) \left(\sum_{k=0}^{m-1} \frac{e^{\alpha_{j+k}[\Lambda(t)-\Lambda(t+s)]}}{\prod_{l=0,l\neq k}^{m} (\alpha_{j+l} - \alpha_{j+k})} + \frac{e^{\alpha_{j+m}[\Lambda(t)-\Lambda(t+s)]}}{\prod_{l=0,l\neq m}^{m} (\alpha_{j+l} - \alpha_{j+m})} \right) \\
&= \left(\prod_{k=0}^{m-1} \alpha_{j+k}\right) \sum_{k=0}^{m} \frac{e^{\alpha_{j+k}[\Lambda(t)-\Lambda(t+s)]}}{\prod_{l=0,l\neq k}^{m} (\alpha_{j+l} - \alpha_{j+k})}
\end{aligned}
$$

□

Remarque 3.2. *La forme analytique* (3.6) *est similaire à celle de l'équation (9) de [Sen et Balakrishnan, 1999], relative au cas particulier de la convolution de lois exponentielles dont les paramètres sont différents deux à deux. Les auteurs présentent ce résultat comme une extension du cas discret de la convolution de distributions géométriques ; leur démonstration fait aussi usage de l'interpolation par un polynôme de Lagrange, à l'instar de la démonstration de la proposition A.1 portée en annexe A.* △

Chapitre 4

L'extension linéaire du processus de Yule

Considérons maintenant, toujours pour un système unique, le cas particulier où l'intensité du processus de naissance non homogène dépend *linéairement* du rang de l'événement. Il s'agit d'une extension directe du processus de Yule, qui sera dénommée *LEYP*, acronyme signifiant *Linear Extension of the Yule Process*. Nous posons donc la :

Définition 4.1. *Un Processus de Yule Linéairement Etendu est défini par l'intensité :*

$$\forall t \in \mathbb{R}_+, \forall j \in \mathbb{N}, \alpha \in \mathbb{R}_+^* \ :$$
$$\mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) = j\right) = (1 + \alpha j)\lambda(t)\mathrm{d}t$$

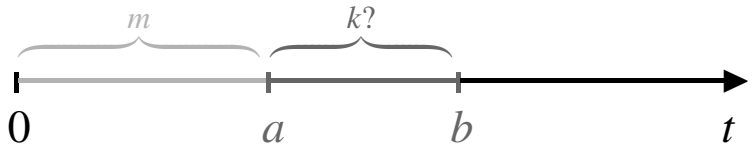
La classe de processus de définition 4.1 contient comme cas particuliers le processus de Poisson non homogène ($\alpha = 0$) et le processus de Yule non homogène ($\alpha = 1$). Si le cas $\alpha = 1$ apparaît trivial, le cas $\alpha = 0$ pose un problème de passage à la limite non évident en ce qui concerne la distribution du processus de comptage $N(t)$, et qui fera l'objet de la sous-section 4.2.

4.1 Distribution conditionnelle du nombre d'événements

Cette section s'intéresse à la distribution conditionnelle du nombre d'événements susceptibles de se produire dans un intervalle donné, dit de *prédiction*, connaissant le nombre d'événements qui se sont produits dans un intervalle antérieur, dit d'*observation*. Nous allons établir cette distribution conditionnelle dans des configurations de plus en plus générales.

4.1.1 Distribution de $N(b) - N(a) \mid N(a-)$

Nous nous intéressons tout d'abord à la configuration la plus simple, illustrée par le schéma suivant :



où l'intervalle d'observation $[0, a]$ commence au début du processus, et l'intervalle de prédiction $[a, b]$ est adjacent à l'intervalle d'observation.

Afin d'alléger la présentation des calculs, nous adoptons dans la suite la notation :

$$\mu(t) = e^{\alpha\Lambda(t)} \tag{4.1}$$

Démontrons d'abord la :

Proposition 4.1. *Dans le cas d'un système obéissant à un LEYP de définition 4.1, le nombre d'événements susceptibles de se produire dans une fenêtre d'observation* $[a, b]$*, sachant que* $N(a-) = m$*, suit une distribution binomiale négative :*

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

Preuve :

Remplaçons dans l'équation (3.6) les α_i par $(1 + i\alpha)$:

$$\begin{aligned}
&\Pr\{N(b) - N(a) = k \mid N(a-) = m\} \\
&= \left(\prod_{j=0}^{k-1}(1 + (m + j)\alpha)\right) \sum_{j=0}^{k} \frac{\exp\{(1 + (m + j)\alpha)\,(\Lambda(a) - \Lambda(b))\}}{\prod_{l=0, l\neq j}^{k}\{(1 + (m + l)\alpha) - (1 + (m + j)\alpha)\}} \\
&= \left(\prod_{j=0}^{k-1}\alpha(\alpha^{-1} + m + j)\right) \sum_{j=0}^{k} \left(\frac{\mu(a)}{\mu(b)}\right)^{\alpha^{-1}+m+j} \left(\prod_{l=0, l\neq j}^{m}\alpha(l - j)\right)^{-1}
\end{aligned}$$

Or :

$$\prod_{j=0}^{k-1}\alpha(\alpha^{-1} + m + j) = \alpha^k \frac{\Gamma(\alpha^{-1} + m + k)}{\Gamma(\alpha^{-1} + m)}$$

et

$$\prod_{l=0, l\neq j}^{k}\alpha(l - j) = \alpha^k(-1)^j j!(k - j)!$$

Donc :

$$\begin{aligned}
&\Pr\{N(b) - N(a) = k \mid N(a-) = m\} \\
&= \frac{\Gamma(\alpha^{-1} + m + k)}{\Gamma(\alpha^{-1} + m)\, k!} \left(\frac{\mu(a)}{\mu(b)}\right)^{\alpha^{-1}+m} \sum_{j=0}^{k} \binom{k}{j} (1)^{k-j} \left(-\frac{\mu(a)}{\mu(b)}\right)^{j}
\end{aligned}$$

D'où en appliquant le théorème binomial :

$$\begin{aligned}
&\Pr\{N(b) - N(a) = k \mid N(a-) = m\} \\
&= \frac{\Gamma(\alpha^{-1} + m + k)}{\Gamma(\alpha^{-1} + m)\, k!} \left(\frac{\mu(a)}{\mu(b)}\right)^{\alpha^{-1}+m} \left(1 - \frac{\mu(a)}{\mu(b)}\right)^{k}
\end{aligned}$$

□

4.1.2 Distribution marginale de $N(b)$

La distribution marginale du processus de comptage $N(b)$ se déduit directement de la proposition 4.1 en posant $a = 0$ et $m = 0$:

Proposition 4.2. *La distribution du processus de comptage $N(b)$ du LEYP est binomiale négative :*

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

Le compensateur du processus de comptage du LEYP est ainsi :

$$\mathrm{E}\left(N(t)\right) = \frac{\mu(t) - 1}{\alpha}$$

Ce résultat est à comparer au NHPP Gamma-mélangé présenté ci-après en sous-section 4.1.3.

4.1.3 Le NHPP Gamma-mélangé

Suivant un résultat exposé par [Lawless, 1987], mais remontant en fait à [Greenwood et Yule, 1920], on obtient aussi une loi binomiale négative avec un NHPP dont la fonction intensité est $\mathrm{E}\left(\mathrm{d}N(t)\right) = v\lambda(t)\mathrm{d}t$, *i.e.* le produit d'une fonction $\lambda(t)$ par un réel strictement positif v, pris comme un facteur aléatoire distribué selon une loi Gamma d'espérance 1 et de variance α : $v \in \mathbb{R}_+ \sim \mathcal{G}(\alpha^{-1}, \alpha)$. Suivant [Greenwood et Yule, 1920], si la v.a. X suit une distribution conditionnelle de Poisson de paramètre aléatoire θ, θ étant une v.a. distribuée selon une loi Gamma de paramètres $\mu, \sigma \in \mathbb{R}_+^*$, alors la distribution marginale de X est binomiale négative $\mathcal{NB}(\mu, 1/(\sigma + 1))$. En effet :

$$\begin{cases} \Pr\left\{X = x \mid \theta\right\} = \dfrac{\theta^x}{x!} e^{-\theta}, \forall x \in \mathbb{N} \\ \theta \in \mathbb{R}_+ \sim \mathcal{G}(\mu, \sigma), \text{ d'où : } f(\theta) = \dfrac{\theta^{\mu-1} e^{-\theta/\sigma}}{\sigma^\mu \Gamma(\mu)} \end{cases}$$

$$\begin{aligned} \Rightarrow \Pr\left\{X = x\right\} &= \int_0^{+\infty} \frac{\theta^x}{x!} e^{-\theta} \frac{\theta^{\mu-1} e^{-\theta/\sigma}}{\sigma^\mu \Gamma(\mu)} \mathrm{d}\theta \\ &= \frac{1}{\Gamma(\mu) x! \sigma^\mu} \int_0^{+\infty} \theta^{x+\mu-1} e^{-\theta(1+\frac{1}{\sigma})} \mathrm{d}\theta \\ &= \frac{\Gamma(\mu + x)}{\Gamma(\mu) x! \sigma^\mu (1 + \frac{1}{\sigma})^{\mu+x}} \\ &= \frac{\Gamma(\mu + x)}{\Gamma(\mu) x!} \left(\frac{1}{\sigma + 1}\right)^\mu \left(\frac{\sigma}{\sigma + 1}\right)^x \end{aligned}$$

L'avant-dernière égalité résulte de la définition et de la propriété suivantes de la fonction Gamma (cf. [Abramowitz et Stegun, 1972]) :

$$\forall x \in \mathbb{R}_+^*, \forall y \in \mathbb{R}_+^*, \quad \Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} \mathrm{d}t = y^x \int_0^{+\infty} t^{x-1} e^{-yt} \mathrm{d}t$$

Si donc $N(t)$ est le processus de comptage du NHPP d'intensité $v\lambda(t)$ avec $v \in \mathbb{R}_+ \sim \mathcal{G}(\alpha^{-1}, \alpha)$, alors $N(t) \sim \mathcal{NB}\left(\alpha^{-1}, (\alpha\Lambda(t) + 1)^{-1}\right)$, où $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$.

4.1.4 La série binomiale puissance

Dans la suite, certaines démonstrations feront usage de la formule de calcul de la série suivante :

$$\forall x \in]0,1[, \forall y \in \mathbb{R}_+, \quad \sum_{j=0}^{\infty} \frac{\Gamma(y+j)}{\Gamma(y)j!} x^j = \frac{1}{(1-x)^y} \tag{4.2}$$

Preuve :

Considérons la variable aléatoire binomiale négative $J \sim \mathcal{NB}(y, 1-x)$ et sommons sa fonction de probabilité définie par (2.4) :

$$\sum_{j=0}^{\infty} \frac{\Gamma(y+j)}{\Gamma(y)j!} (1-x)^y x^j = 1$$

D'où :

$$\sum_{j=0}^{\infty} \frac{\Gamma(y+j)}{\Gamma(y)j!} x^j = \frac{1}{(1-x)^y}$$

□

4.1.5 Distribution marginale de $N(b) - N(a)$

La distribution marginale binomiale négative de $N(b)-N(a)$ est une conséquence importante de la proposition 4.1 :

Proposition 4.3.

$$\forall a, b \in \mathbb{R}_+, \text{ tels que } a < b, \quad N(b) - N(a) \sim \mathcal{NB}\left(\alpha^{-1}, \frac{1}{\mu(b) - \mu(a) + 1}\right)$$

Preuve :

Calculons l'espérance de $\Pr\{N(b) - N(a) = k \mid N(a-) = m\}$, puis utilisons (4.2) :

$$\begin{aligned}
&\Pr\{N(b) - N(a) = k\} \\
&= \sum_{m=0}^{\infty} \Pr\{N(b) - N(a) = k \mid N(a-) = m\} \Pr\{N(a-) = m\} \\
&= \sum_{m=0}^{\infty} \frac{\Gamma(\alpha^{-1} + m + k)}{\Gamma(\alpha^{-1} + m)\, k!} \left(\frac{\mu(a)}{\mu(b)}\right)^{\alpha^{-1}+m} \left(\frac{\mu(b) - \mu(a)}{\mu(b)}\right)^k \frac{\Gamma(\alpha^{-1} + m)}{\Gamma(\alpha^{-1})\, m!} \left(\frac{1}{\mu(a)}\right)^{\alpha^{-1}} \left(\frac{\mu(a) - 1}{\mu(a)}\right)^m \\
&= \frac{\Gamma(\alpha^{-1} + k)}{\Gamma(\alpha^{-1})\, k!} \left(\frac{1}{\mu(b)}\right)^{\alpha^{-1}} \left(\frac{\mu(b) - \mu(a)}{\mu(b)}\right)^k \sum_{m=0}^{\infty} \frac{\Gamma(\alpha^{-1} + m + k)}{\Gamma(\alpha^{-1} + k)\, m!} \left(\frac{\mu(a) - 1}{\mu(b)}\right)^m
\end{aligned}$$

$$= \frac{\Gamma(\alpha^{-1}+k)}{\Gamma(\alpha^{-1})\,k!}\left(\frac{1}{\mu(b)}\right)^{\alpha^{-1}}\left(\frac{\mu(b)-\mu(a)}{\mu(b)}\right)^{k}\left(1-\frac{\mu(a)-1}{\mu(b)}\right)^{-(\alpha^{-1}+k)}$$
$$= \frac{\Gamma(\alpha^{-1}+k)}{\Gamma(\alpha^{-1})\,k!}\left(\frac{1}{\mu(b)-\mu(a)+1}\right)^{\alpha^{-1}}\left(\frac{\mu(b)-\mu(a)}{\mu(b)-\mu(a)+1}\right)^{k}$$ □

4.1.6 Distribution de $N(a-) \mid N(b) - N(a)$

Le résultat suivant nous sera utile par la suite :

Proposition 4.4. *La probabilité conditionnelle de $N(a-)$, connaissant $N(b) - N(a)$ ($0 < a < b$) est binomiale négative :*

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

Preuve :

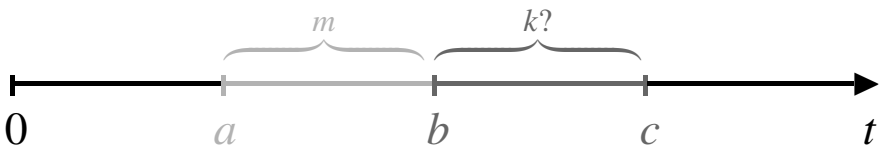
$$\begin{aligned}
&\Pr\{N(a-) = j \mid N(b) - N(a) = m\} \\
&= \frac{\Pr\{N(a-) = j, N(b) - N(a) = m\}}{\Pr\{N(b) - N(a) = m\}} \\
&= \frac{\Pr\{N(b) - N(a) = m \mid N(a-) = j\}\Pr\{N(a-) = j\}}{\Pr\{N(b) - N(a) = m\}}
\end{aligned}$$

Utilisons les propositions 4.2 et 4.3 :

$$\begin{aligned}
&= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+m)\,k!}\left(\frac{\mu(a)}{\mu(b)}\right)^{\alpha^{-1}+m}\left(\frac{\mu(b)-\mu(a)}{\mu(b)}\right)^{k}\frac{\Gamma(\alpha^{-1}+m)}{\Gamma(\alpha^{-1})\,m!}\left(\frac{1}{\mu(a)}\right)^{\alpha^{-1}}\left(\frac{\mu(a)-1}{\mu(a)}\right)^{m}\times \\
&\quad\times\left(\frac{\Gamma(\alpha^{-1}+k)}{\Gamma(\alpha^{-1})\,k!}\left(\frac{1}{\mu(b)-\mu(a)+1}\right)^{\alpha^{-1}}\left(\frac{\mu(b)-\mu(a)}{\mu(b)-\mu(a)+1}\right)^{k}\right)^{-1} \\
&= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+k)\,m!}\left(\frac{\mu(b)-\mu(a)+1}{\mu(b)}\right)^{\alpha^{-1}+k}\left(\frac{\mu(a)-1}{\mu(b)}\right)^{m}
\end{aligned}$$ □

4.1.7 Distribution de $N(c) - N(b) \mid N(b-) - N(a)$

La première généralisation de la proposition 4.1 considère un intervalle d'observation $[a, b]$ ne commençant pas nécessairement au début du processus, et un intervalle de prédiction $[b, c]$ adjacent à l'intervalle d'observation, comme illustré ci-dessous :



Démontrons la :

Proposition 4.5. *La probabilité conditionnelle de $N(c) - N(b)$, connaissant $N(b-) - N(a)$ ($0 < a < b < c$) est binomiale négative :*

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

Preuve :

Selon la formule des probabilités totales :

$$\begin{aligned}
&\Pr\{N(c) - N(b) = k \mid N(b-) - N(a) = m\} \\
&= \sum_{j=0}^{\infty} \Pr\{N(c) - N(b) = k \mid N(b-) - N(a) = m, N(a-) = j\} \times \\
&\qquad \times \Pr\{N(a-) = j \mid N(b-) - N(a) = m\} \\
&= \sum_{j=0}^{\infty} \Pr\{N(c) - N(b) = k \mid N(b-) = j + m\} \Pr\{N(a-) = j \mid N(b-) - N(a) = m\}
\end{aligned}$$

Utilisant la proposition 4.4 :

$$\begin{aligned}
&\Pr\{N(a-) = j \mid N(b-) - N(a) = m\} \\
&= \frac{\Gamma(\alpha^{-1} + j + m)}{\Gamma(\alpha^{-1} + m)j!} (\mu(b) - \mu(a) + 1)^{\alpha^{-1}+m} (\mu(a) - 1)^j / \mu(b)^{\alpha^{-1}+m+j}
\end{aligned}$$

Et donc :

$$\begin{aligned}
&\Pr\{N(c) - N(b) = k \mid N(b-) - N(a) = m\} \\
&= \sum_{j=0}^{\infty} \frac{\Gamma(\alpha^{-1} + j + m + k)}{\Gamma(\alpha^{-1} + j + m)k!} (\mu(b)/\mu(c))^{\alpha^{-1}+j+m} (1 - \mu(b)/\mu(c))^k \times \\
&\qquad \times \frac{\Gamma(\alpha^{-1} + j + m)}{\Gamma(\alpha^{-1} + m)j!} (\mu(b) - \mu(a) + 1)^{\alpha^{-1}+m} (\mu(a) - 1)^j / \mu(b)^{\alpha^{-1}+m+j} \\
&= \frac{\Gamma(\alpha^{-1} + m + k)}{\Gamma(\alpha^{-1} + m)k!} (\mu(b)/\mu(c))^{\alpha^{-1}+m} (1 - \mu(b)/\mu(c))^k (\mu(b) - \mu(a) + 1)^{\alpha^{-1}+m} / \mu(b)^{\alpha^{-1}+m} \times \\
&\quad \times \sum_{j=0}^{\infty} \frac{\Gamma(\alpha^{-1} + m + k + j)}{\Gamma(\alpha^{-1} + m + k)j!} \left[(\mu(b)/\mu(c))(\mu(a) - 1)/\mu(b)\right]^j
\end{aligned}$$

Utilisons à nouveau la série (4.2) :

$$= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+m)k!}\left(1/\mu(c)\right)^{\alpha^{-1}+m}\left(1-\mu(b)/\mu(c)\right)^{k}\left(\mu(b)-\mu(a)+1\right)^{\alpha^{-1}+m}\left(1-\frac{\mu(a)-1}{\mu(c)}\right)^{-(\alpha^{-1}+m+k)}$$
$$= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+m)k!}\frac{\left(\mu(b)-\mu(a)+1\right)^{\alpha^{-1}+m}\left(\mu(c)-\mu(b)\right)^{k}}{\left(\mu(c)-\mu(a)+1\right)^{\alpha^{-1}+m+k}}$$

□

4.1.8 Distribution de $N(b-)-N(a) \mid N(c)-N(b)$

La généralisation suivante de la proposition 4.4 nous sera utile plus tard :

Proposition 4.6. *Pour un LEYP de définition 4.1, la probabilité conditionnelle de $N(b-)-N(a)$, connaissant $N(c)-N(b)$, avec $0<a<b<c$, est binomiale négative :*

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

Preuve :

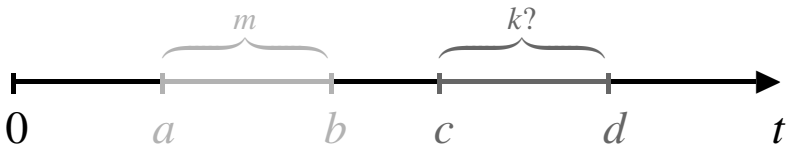
Procédons comme pour la démonstration de la proposition 4.4 :

$$\begin{aligned}
&\Pr\{N(b-)-N(a)=m \mid N(c)-N(b)=k\} \\
&= \frac{\Pr\{N(c)-N(b)=k \mid N(b-)-N(a)=m\}\Pr\{N(b-)-N(a)=m\}}{\Pr\{N(c)-N(b)=k\}} \\
&= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+m)k!}\left(\frac{\mu(b)-\mu(a)+1}{\mu(c)-\mu(a)+1}\right)^{\alpha^{-1}+m}\left(\frac{\mu(c)-\mu(b)}{\mu(c)-\mu(a)+1}\right)^{k}\times \\
&\quad\times\frac{\Gamma(\alpha^{-1}+m)}{\Gamma(\alpha^{-1})m!}\left(\frac{1}{\mu(b)-\mu(a)+1}\right)^{\alpha^{-1}}\left(\frac{\mu(b)-\mu(a)}{\mu(b)-\mu(a)+1}\right)^{m}\times \\
&\quad\times\left(\frac{\Gamma(\alpha^{-1}+k)}{\Gamma(\alpha^{-1})k!}\left(\frac{1}{\mu(c)-\mu(b)+1}\right)^{\alpha^{-1}}\left(\frac{\mu(c)-\mu(b)}{\mu(c)-\mu(b)+1}\right)^{k}\right)^{-1} \\
&= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+k)m!}\left(\frac{\mu(c)-\mu(b)+1}{\mu(c)-\mu(a)+1}\right)^{\alpha^{-1}+k}\left(\frac{\mu(b)-\mu(a)}{\mu(c)-\mu(a)+1}\right)^{m}
\end{aligned}$$

□

4.1.9 Distribution de $N(d) - N(c) \mid N(b) - N(a)$

La proposition suivante fournit un résultat plus général, en ce que l'intervalle de prédiction $[c, d]$ est disjoint de l'intervalle d'observation $[a, b]$:



Proposition 4.7. *Soit un système obéissant à un LEYP défini par* (4.1)*, et les instants* $0 < a < b < c < d$*. La probabilité conditionnelle de* $N(d) - N(c)$*, sachant que* $N(b) - N(a) = m$*, est binomiale négative :*

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

Preuve :

Selon la formule des probabilités totales, et en appliquant la proposition 4.5 :

$$\begin{aligned}
&\Pr\{N(d) - N(c) = k \mid N(b) - N(a) = m\} \\
&= \sum_{j=0}^{\infty} \Pr\{N(d) - N(c) = k \mid N(b) - N(a) = m, N(c-) - N(b+) = j\} \times \\
&\qquad \times \Pr\{N(c-) - N(b+) = j \mid N(b) - N(a) = m\} \\
&= \sum_{j=0}^{\infty} \Pr\{N(d) - N(c) = k \mid N(c-) - N(a) = m + j\} \times \\
&\qquad \times \Pr\{N(c-) - N(b+) = j \mid N(b) - N(a) = m\} \\
&= \sum_{j=0}^{\infty} \frac{\Gamma(\alpha^{-1} + m + j + k)}{\Gamma(\alpha^{-1} + m + j)k!} \left(\frac{\mu(c) - \mu(a) + 1}{\mu(d) - \mu(a) + 1}\right)^{\alpha^{-1}+m+j} \left(\frac{\mu(d) - \mu(c)}{\mu(d) - \mu(a) + 1}\right)^{k} \times \\
&\qquad \times \frac{\Gamma(\alpha^{-1} + m + j)}{\Gamma(\alpha^{-1} + m)j!} \left(\frac{\mu(b) - \mu(a) + 1}{\mu(c) - \mu(a) + 1}\right)^{\alpha^{-1}+m} \left(\frac{\mu(c) - \mu(b)}{\mu(c) - \mu(a) + 1}\right)^{j} \\
&= \frac{\Gamma(\alpha^{-1} + m + k)}{\Gamma(\alpha^{-1} + m)k!} \left[\left(\frac{\mu(c) - \mu(a) + 1}{\mu(d) - \mu(a) + 1}\right)\left(\frac{\mu(b) - \mu(a) + 1}{\mu(c) - \mu(a) + 1}\right)\right]^{\alpha^{-1}+m} \left(\frac{\mu(d) - \mu(c)}{\mu(d) - \mu(a) + 1}\right)^{k} \times \\
&\quad \times \sum_{j=0}^{\infty} \frac{\Gamma(\alpha^{-1} + m + j + k)}{\Gamma(\alpha^{-1} + m + k)j!} \left[\left(\frac{\mu(c) - \mu(a) + 1}{\mu(d) - \mu(a) + 1}\right)\left(\frac{\mu(c) - \mu(b)}{\mu(c) - \mu(a) + 1}\right)\right]^{j}
\end{aligned}$$

utilisant la série (4.2) :

$$= \frac{\Gamma(\alpha^{-1} + m + k)}{\Gamma(\alpha^{-1} + m)k!} \left(\frac{\mu(b) - \mu(a) + 1}{\mu(d) - \mu(a) + 1}\right)^{\alpha^{-1}+m} \left(\frac{\mu(d) - \mu(c)}{\mu(d) - \mu(a) + 1}\right)^{k} \left(1 - \frac{\mu(c) - \mu(b)}{\mu(d) - \mu(a) + 1}\right)^{-(\alpha^{-1}+m+k)}$$

$$= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+m)k!}\frac{(\mu(b)-\mu(a)+1)^{\alpha^{-1}+m}(\mu(d)-\mu(c))^k}{(\mu(d)-\mu(c)+\mu(b)-\mu(a)+1)^{\alpha^{-1}+m+k}} \qquad \square$$

Ce résultat est important en ce qu'il fournit une forme analytique explicite de la distribution du nombre possible d'événements dans un intervalle de temps où une prédiction est souhaitée, connaissant le nombre d'événements observés dans un intervalle de temps antérieur, non nécessairement adjacent. Tous les calculs de prédiction aux fins de validation du modèle LEYP sur un jeu de données réel, ou de prévision de casses pour la prise de décision en matière de réhabilitation, s'appuient sur cette formule.

4.2 Distribution limite pour α tendant vers 0+

Nous pouvons montrer que la distribution du processus de comptage du LEYP tend vers celle d'un NHPP lorsque α tend vers 0. Cela revient à montrer la :

Proposition 4.8. *La distribution binomiale négative $\mathcal{NB}(\alpha^{-1}, e^{-\alpha\Lambda(t)})$ tend vers la distribution de Poisson $\mathcal{P}o(\Lambda(t))$ lorsque α tend vers* 0*.*

Preuve :

Remarquons tout d'abord que :

$$\begin{aligned}\alpha^m\Gamma(\alpha^{-1}+m)/\Gamma(\alpha^{-1}) &= \alpha^m\frac{1}{\alpha}\left(\frac{1}{\alpha}+1\right)\left(\frac{1}{\alpha}+2\right)\cdots\left(\frac{1}{\alpha}+m-1\right)\\ &= \frac{\alpha}{\alpha}\left(\frac{\alpha}{\alpha}+\alpha\right)\left(\frac{\alpha}{\alpha}+2\alpha\right)\cdots\left(\frac{\alpha}{\alpha}+(m-1)\alpha\right)\\ &= (1+\alpha)(1+2\alpha)\cdots(1+(m-1)\alpha)\end{aligned}$$

D'où :

$$\lim_{\alpha\to 0+}\alpha^m\Gamma(\alpha^{-1}+m)/\Gamma(\alpha^{-1}) = 1$$

Nous avons ensuite par définition :

$$\forall x\in\mathbb{R},\quad e^x = \sum_{j=0}^{\infty} x^j/j!$$

Il s'ensuit que :

$$\begin{aligned}1-e^{-\alpha\Lambda(t)} &= 1-\left(1+[-\alpha\Lambda(t)]+[-\alpha\Lambda(t)]^2/2!+[-\alpha\Lambda(t)]^3/3!+\cdots\right)\\ &= \alpha\Lambda(t)-[\alpha\Lambda(t)]^2/2!+[\alpha\Lambda(t)]^3/3!-\cdots\end{aligned}$$

D'où :

$$\frac{1 - e^{-\alpha\Lambda(t)}}{\alpha} = \Lambda(t) - \alpha\Lambda(t)^2/2! + \alpha^2\Lambda(t)^3/3! - \cdots$$

Et donc :

$$\lim_{\alpha\to 0+} \frac{1 - e^{-\alpha\Lambda(t)}}{\alpha} = \Lambda(t)$$

Appliquons cela à la fonction de probabilité de la distribution $\mathcal{NB}(\alpha^{-1}, e^{-\alpha\Lambda(t)})$:

$$\begin{aligned}
\lim_{\alpha\to 0+} \Pr\{N(t) = m\} &= \lim_{\alpha\to 0+} \frac{\Gamma(\alpha^{-1} + m)}{\Gamma(\alpha^{-1})m!} \left(e^{-\alpha\Lambda(t)}\right)^{\alpha^{-1}} \left(1 - e^{-\alpha\Lambda(t)}\right)^m \\
&= \lim_{\alpha\to 0+} \frac{\Gamma(\alpha^{-1} + m)}{\Gamma(\alpha^{-1})} \alpha^m \frac{e^{-\Lambda(t)}}{m!} \left(\frac{1 - e^{-\alpha\Lambda(t)}}{\alpha}\right)^m \\
&= \frac{\Lambda(t)^m}{m!} e^{-\Lambda(t)}
\end{aligned}$$

Qui est bien la fonction de probabilité de la distribution $\mathcal{P}o(\Lambda(t))$. □

Chapitre 5

Inférence sur une fenêtre d'observation

Nous abordons maintenant le problème de l'estimation des paramètres du LEYP à partir d'un échantillon de systèmes observés chacun au sein d'une fenêtre d'observation. La motivation pratique tient au fait que le LEYP doit pouvoir modéliser les occurrences de défaillances survenant sur des systèmes dont l'historique des défaillances n'a pas été archivé à compter de leur mise en service. Tel est le cas des canalisations d'eau potable, massivement posées dans les centres urbains européens avant 1940, et dont les archives de maintenance ne sont accessibles qu'après 1985 au plus tôt.

5.1 Vraisemblance du LEYP

Considérons un objet technique observé dans $[a, b]$, avec $a \in \mathbb{R}_+$, $b \in \mathbb{R}_+^*$, et $a < b$, et subissant $m \in \mathbb{N}$ défaillances dans $[a, b]$, aux instants $t_1 < \ldots < t_j < \ldots < t_m$. La méthode d'inférence des paramètres du LEYP à partir des observations consiste à construire la fonction de vraisemblance des paramètres connaissant les t_j, puis à rechercher la valeur des paramètres qui maximise cette fonction. La construction intuitive de la fonction de vraisemblance, telle que présentée dans [Samset, 1994] dans le cas du NHPP, consiste à faire le produit des probabilités de n'observer aucun événement sur les intervalles $]t_j, t_{j+1}[$, et des limites quand $h \to 0+$ des probabilités d'observer un événement sur $[t_j, t_j + h[$, toutes ces probabilités étant conditionnelles connaissant $N(t_j)$. La difficulté provient de ce que le système n'est pas observé sur $[0, a[$, et que donc $N(a)$ est inconnu, ainsi que l'ordre des défaillances observées successives ; l'information collectée au sein d'une fenêtre d'observation est dite ainsi tronquée à gauche.

Afin de construire la fonction de vraisemblance de façon rigoureuse, nous nous appuyons sur le concept général de vraisemblance au sens de Jacod exposé par [Andersen *et al.*, 1993]. La vraisemblance du processus théorique LEYP de vecteur de paramètres $\boldsymbol{\theta}$ connaissant une séquence de défaillances est formellement définie comme le produit intégral :

$$L(\boldsymbol{\theta}) = \prod_{t \in [a,b]} \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)^{\Delta N(t)} \left(1 - \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)\right)^{1-\Delta N(t)} \tag{5.1}$$

$$\text{où :} \quad \Delta N(t) = N(t) - N(t-)$$

L'utilisation du produit intégral pour définir la vraisemblance d'un processus de comptage revient à considérer la fenêtre d'observation $[a, b]$ comme la succession d'un nombre infini

dénombrable d'intervalles de longueur infinitésimalle, chacun subissant une expérience de Bernouilli, dont le résultat, indépendant des expériences qui précèdent, est soit une défaillance avec la probabilité $\mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)$, soit une absence de défaillance avec la probabilité complémentaire $1 - \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)$.

Afin d'être en mesure de donner à l'équation (5.1) une forme analytique explicite, nous devons au préalable établir la :

Proposition 5.1.

$$\begin{aligned}
&\mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right) \\
&= \Big(\alpha^{-1} + (N(t-) - N(a))\Big)\mathrm{d}\ln\left(\mu(t) - \mu(a) + 1\right) \\
&= \Big(1 + (N(t-) - N(a))\alpha\Big)\frac{\mu(t)\lambda(t)\mathrm{d}t}{\mu(t) - \mu(a) + 1}
\end{aligned}$$

Preuve :

En conséquence de la définition 4.1, et utilisant la proposition 4.5 et la continuité de la fonction $\mu(.)$:

$$\begin{aligned}
&\mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right) = \Pr\left\{N(t + \mathrm{d}t) - N(t) = 1 \mid N(t-) - N(a)\right\} \\
&= \Big(\alpha^{-1} + N(t-) - N(a)\Big)\left(\frac{\mu(t) - \mu(a) + 1}{\mu(t + \mathrm{d}t) - \mu(a) + 1}\right)^{\alpha^{-1} + N(t-) - N(a)}\left(\frac{\mu(t + \mathrm{d}t) - \mu(t)}{\mu(t + \mathrm{d}t) - \mu(a) + 1}\right) \\
&= \Big(\alpha^{-1} + N(t-) - N(a)\Big)\frac{\mathrm{d}\mu(t)}{\mu(t) - \mu(a) + 1} \\
&= \Big(\alpha^{-1} + (N(t-) - N(a))\Big)\mathrm{d}\ln\left(\mu(t) - \mu(a) + 1\right)
\end{aligned}$$

(On peut alternativement expliciter $\mathrm{d}\mu(t)$ en $\alpha\mu(t)\lambda(t)\mathrm{d}t$.) □

Nous pouvons maintenant montrer la :

Proposition 5.2. *La vraisemblance du processus théorique LEYP de vecteur de paramètres $\boldsymbol{\theta}$ connaissant une séquence de défaillances est :*

$$L(\boldsymbol{\theta}) = \alpha^m \frac{\Gamma(\alpha^{-1} + m)}{\Gamma(\alpha^{-1})} \frac{\prod_{j=1}^{m} \mu(t_j)\lambda(t_j)}{(\mu(b) - \mu(a) + 1)^{\alpha^{-1} + m}}$$

Preuve :

Du fait du nombre fini m de sauts dans $[a, b]$:

$$\prod_{t \in [a,b]} \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)^{\Delta N(t)} \left(1 - \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)\right)^{1 - \Delta N(t)}$$

$$= \prod_{j=1}^{m} \mathrm{E}\left(\mathrm{d}N(t_j) \mid \mathscr{N}_{[a,t_j[}\right) \prod_{j=0}^{m} \prod_{t\in]t_j,t_{j+1}[} \left(1 - \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)\right)$$

Ignorant $(\mathrm{d}t)^m$ (*cf.* [Andersen *et al.*, 1993]), le facteur de gauche s'écrit :

$$\begin{aligned}
&\prod_{j=1}^{m} \mathrm{E}\left(\mathrm{d}N(t_j) \mid N(t_j-) - N(a) = j-1\right) \\
&= \prod_{j=1}^{m} (1 + (j-1)\alpha) \frac{\mu(t_j)\lambda(t_j)}{\mu(t_j) - \mu(a) + 1} = \alpha^m \frac{\Gamma(\alpha^{-1} + m)}{\Gamma(\alpha^{-1})} \prod_{j=1}^{m} \frac{\mu(t_j)\lambda(t_j)}{\mu(t_j) - \mu(a) + 1}
\end{aligned}$$

En utilisant la propriété du produit intégral $\prod (1 - \mathrm{d}X) = \exp\left(-\int \mathrm{d}X\right)$ (*cf.* [Andersen *et al.*, 1993]), le facteur de droite s'écrit :

$$\begin{aligned}
&\prod_{j=0}^{m} \prod_{t\in]t_j,t_{j+1}[} \left(1 - \mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) - N(a) = j\right)\right) \\
&= \prod_{j=0}^{m} \exp\left(-\int_{t_j}^{t_{j+1}} \left(\alpha^{-1} + j\right) \frac{\mathrm{d}\mu(t)}{\mu(t) - \mu(a) + 1}\right) \\
&= \prod_{j=0}^{m} \left(\frac{\mu(t_j) - \mu(a) + 1}{\mu(t_{j+1}) - \mu(a) + 1}\right)^{\alpha^{-1}+j} = \frac{\prod_{j=1}^{m} \mu(t_j) - \mu(a) + 1}{(\mu(b) - \mu(a) + 1)^{\alpha^{-1}+m}}
\end{aligned}$$

La simplification par $\prod_{j=1}^{m} \mu(t_j) - \mu(a) + 1$ achève la démonstration. □

Pour un n-échantillon *aléatoire* d'objets techniques supposés défaillir indépendamment les uns des autres m_i fois dans $[a_i, b_i], i \in \{1, \ldots, n\}$, la vraisemblance globale est simplement le produit des vraisemblances individuelles :

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta})$$

En pratique, on considère plutôt la log-vraisemblance qui implique des calculs de sommations, toujours numériquement préférables à des produits. La log-vraisemblance pour n objets techniques s'écrit :

$$\begin{aligned}
\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Bigg(& m_i \ln \alpha + \ln \Gamma\left(\alpha^{-1} + m_i\right) - \ln \Gamma\left(\alpha^{-1}\right) \\
& - \left(\alpha^{-1} + m_i\right) \ln\left(\mu(b_i) - \mu(a_i) + 1\right) + \sum_{j=1}^{m_i} \ln \lambda(t_{ij}) + \alpha \Lambda(t_{ij}) \Bigg)
\end{aligned} \tag{5.2}$$

5.2 Estimation des paramètres du LEYP

5.2.1 Estimateur du maximum de vraisemblance

Comme dans nombre de problèmes d'estimation des paramètres de modèles statistiques, il est ici pratique de procéder en maximisant le logarithme népérien de la vraisemblance fourni par l'équation (5.2). Le schéma théorique de l'estimation du maximum de vraisemblance consiste dans notre cas à considérer une population *infinie* d'objets techniques dont l'occurrence des défaillances obéit à un LEYP d'intensité de Yule-Weibull-Cox de vecteur de paramètres $\boldsymbol{\vartheta}$. Il convient ici de préciser que le vecteur *théorique* $\boldsymbol{\vartheta}$ est *non aléatoire*, sa valeur étant fixe bien qu'inconnue. L'estimateur du maximum de vraisemblance de $\boldsymbol{\vartheta}$, connaissant les séquences de défaillances de n canalisations, est un vecteur aléatoire :

$$\hat{\boldsymbol{\theta}}_n = \arg_{\boldsymbol{\theta}} \max \ln L(\boldsymbol{\theta})$$

dont nous conjecturons qu'il est asymptotiquement (i) sans biais, (ii) efficace et (iii) normalement distribué :

$$\text{(i)} \quad \lim_{n\to\infty} \mathrm{E}\left(\hat{\boldsymbol{\theta}}_n\right) = \boldsymbol{\vartheta}$$

$$\text{(ii)} \quad \lim_{n\to\infty} \mathrm{Var}\left(\hat{\boldsymbol{\theta}}_n\right) = \left(-\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta^2}\right)^{-1}_{\boldsymbol{\vartheta}}$$

$$\text{(iii)} \quad \lim_{n\to\infty} \left(-\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta^2}\right)^{\frac{1}{2}}_{\boldsymbol{\vartheta}} \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\vartheta}\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

où $\mathbf{0}$ et $\mathbf{1}$ sont respectivement le vecteur nul et la matrice identité de même dimension que $\boldsymbol{\theta}$.

La démonstration de ces propriétés reste à établir dans le cas du LEYP, et devrait idéalement suivre le schéma exposé dans les ouvrages classiques de statistique mathématique traitant de l'estimation du maximum de vraisemblance, en particulier [Rao, 1973] et [Cox et Hinkley, 1974]. Cependant, les propriétés asymptotiques de l'estimateur du maximum de vraisemblance reposent sur des conditions de régularité de la fonction de vraisemblance qui nous paraissent difficiles à vérifier avec la forme analytique (5.2). Nous nous contenterons donc de vérifier graphiquement, avec un jeu de données simulées de plusieurs milliers d'individus avec des taux de défaillance du même ordre de grandeur que ceux observés dans les applications du modèle, que la fonction de log-vraisemblance a une forme qui autorise la recherche de son maximum ; cette étude par simulations numériques a été menée dans le cadre du modèle plus général ζ-LEYP, présenté au chapitre 6, et sa présentation se situe en fin de ce chapitre.

5.2.2 Test d'hypothèse nulle sur les paramètres

La propriété (ii) énoncée ci-avant à la sous-section 5.2.1 nous permet d'estimer la matrice des variances-covariances d'un vecteur de paramètres estimés $\hat{\boldsymbol{\theta}}$:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = \left(-\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta^2}\right)^{-1}_{\hat{\boldsymbol{\theta}}} \tag{5.3}$$

Nous pouvons ainsi implémenter un test *d'hypothèse nulle* sur les paramètres : le but est d'inférer si la valeur estimée d'un paramètre diffère *significativement* ou non (*i.e.* relativement à un risque d'erreur arbitraire) d'une valeur caractéristique. Nous avons choisi d'implémenter cette procédure inférentielle sous la forme, fréquemment utilisée en pratique (par exemple, dans les procédures du système SAS), du test dit *de Wald* décrit dans [Greenwood et Nikulin, 1996] ; ce test repose sur la distribution du χ^2 à un degré de liberté que suit le carré d'une variable normale centrée réduite (*i.e.* d'espérance nulle et de variance unité).

Ainsi, pour les β associés aux covariables d'un modèle LEYP d'intensité de Yule-Weibull-Cox, il est pertinent de tester par rapport à 0, *i.e.* l'absence d'effet des covariables. La statistique du test se calcule simplement comme le carré de l'écart de la valeur estimatée du paramètre $\hat{\theta}_k$ à la valeur caractéristique θ_{k0}, rapporté à la variance d'estimation :

$$\frac{\left(\hat{\theta}_k - \theta_{k0}\right)^2}{\widehat{\mathrm{Var}}(\hat{\theta}_k)} \sim \chi^2_1$$

où $\widehat{\mathrm{Var}}(\hat{\theta}_k)$ est le $k^{\text{ième}}$ terme diagonal de la matrice $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ définie par l'équation (5.3).

Dans le cas de α et δ, paramètres contraints ($\alpha > 0$ et $\delta \geqslant 1$), l'écart-type d'estimation peut être suffisamment important pour invalider l'hypothèse de normalité asymptotique. Il semble donc plus rigoureux, ou en tout cas moins problématique à la mise en oeuvre, de tester α et δ par le test dit « du rapport de vraisemblance » ; cela nécessite pour α d'estimer sur le même jeu de données un modèle alternatif avec $\alpha = 0$, *i.e.* un NHPP, de calculer la statistique *LR* du double de la différence des log-vraisemblances :

$$LR = 2\left(\ln L_{\mathrm{LEYP}} - \ln L_{\mathrm{NHPP}}\right)$$

et de considérer que, sous l'hypothèse nulle $\alpha = 0$, la statistique *LR* suit une distribution du χ^2, dont le degré de liberté est égal à la différence entre le nombre de paramètres du modèle « complet », ici le LEYP, et le nombre de paramètres du modèle « réduit », ici le NHPP. La procédure de test est analogue pour δ avec l'hypothèse nulle d'absence de vieillissement $\delta = 1$.

5.2.3 Algorithme d'estimation des paramètres

En pratique, nous avons, au début de notre travail de recherche, utilisé l'algorithme *classique* d'optimisation dit de *Levenberg-Marquardt* tel qu'il est décrit par [Bard, 1974] et [Fletcher, 1987]. Cet algorithme consiste à annuler le vecteur gradient (dérivées partielles premières de la log-vraisemblance par rapport aux paramètres), en calculant l'intersection avec l'axe des abscisses d'une approximation linéaire de ce dernier ; cette approximation linéaire s'obtient par la formule de Taylor à l'ordre 1, et nécessite donc de dériver à leur tour les composantes du gradient, donc de calculer les composantes de la matrice hessienne des dérivées partielles secondes de la log-vraisemblance par rapport aux paramètres. Lorsqu'en outre la matrice hessienne n'est pas définie positive (*i.e.* a au moins une valeur propre négative) tous ses termes diagonaux sont itérativement incrémentés d'une quantité positive suffisamment élevée jusqu'à ce que la hessienne ainsi transformée soit définie positive ; lorsque le vecteur $\boldsymbol{\theta}^{(.)}$ exploré à l'itération courante produit une log-vraisemblance supérieure à celle de l'itération précédente, la diagonale de la hessienne est décrémentée. Nous avons choisi d'utiliser les formes analytiques explicites

des composantes du gradient et de la hessienne, plutôt que d'implémenter un algorithme les calculant de façon approchée (par exemple par différences finies à droite) qui alourdirait le temps de calcul par des appels trop nombreux à la fonction calculant la log-vraisemblance. Les formes analytiques explicites des dérivées partielles sont portées en annexe B ; ces formules sont relatives à un seul objet technique par souci de simplicité d'écriture.

L'algorithme de Levenberg-Marquardt n'est cependant pas conçu pour traiter les problèmes d'optimisation sous contrainte ; or le modèle LEYP impose de borner deux de ses paramètres : $\alpha > 0$ et $\delta \geqslant 1$. Il a donc été tenté dans un premier temps d'effectuer une transformation des paramètres : $\alpha = \exp(\alpha_0)$ et $\delta = 1 + \exp(\delta_0)$. Remplacer α et δ par α_0 et δ_0 dans le problème d'optimisation permet effectivement d'obtenir un problème sans contrainte ; la contrepartie est que la fonction de log-vraisemblance n'est pas convexe en α_0 et δ_0, et présente un point d'inflexion en deçà de l'optimum et un plateau horizontal pour $\alpha_0 \to -\infty$ et $\delta_0 \to -\infty$. Cela semble susceptible de poser des problèmes de convergence, et fréquemment de faire tendre δ vers 1 lorsque le vieillissement est peu marqué.

La solution adoptée plus récemment repose sur l'algorithme dit de *Nelder-Mead*, bien décrit par [Press *et al.*, 2002]. Pour un modèle à p paramètres, cet algorithme explore l'espace des paramètres grâce à un polyhèdre à $p + 1$ sommets (appelé aussi parfois *simplex*), alternativement étiré dans une direction, contracté dans une ou plusieurs directions, ou déformé par réflexion d'un sommet par rapport à l'hyperplan des p autres sommets, de façon à finir par se contracter sur l'optimun ; le processus d'optimisation stoppe lorsque la différence entre les valeurs supérieure et inférieure prises par la fonction objectif sur l'ensemble des sommets est en dessous d'un seuil fixé. Les bornes sur α et δ sont simplement prises en compte en interdisant aux déformations du polyhèdre de produire des valeurs de α inférieures à un réel positif proche de 0 (10^{-4} en pratique), ainsi que des valeurs de δ strictement inférieures à 1.

L'algorithme de Nelder-Mead a aussi l'avantage de ne nécessiter aucun calcul de dérivées de la fonction objectif par rapport aux paramètres ; la hessienne est cependant estimée par calcul approché à l'optimum afin d'estimer la matrice de variances-covariances des estimations. Notre expérience d'estimations de modèles avec des jeux de données artificiels, donc de paramètres connus, en faisant varier la valeur initiale du vecteur de paramètres n'a pas mis à ce jour en évidence de problème de convergence ou d'estimation erronée avec l'algorithme de Nelder-Mead. Cet algorithme est sans doute un peu plus lent que celui de Levenberg-Marquardt, sans que cela soit pénalisant en pratique.

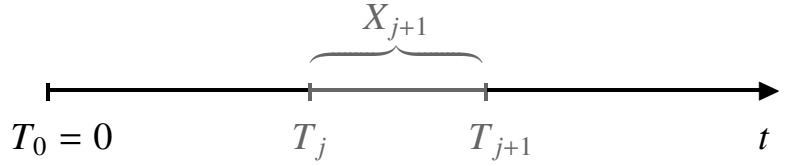
5.3 Validation de la procédure d'estimation du modèle LEYP

Estimer les paramètres du modèle LEYP sur un jeu de données implique la mise en oeuvre d'un programme informatique dont la complexité d'écriture est source potentielle de nombreuses erreurs, de nature tant algorithmique que numérique. Afin de s'assurer que la procédure que nous avons écrite est satisfaisante, nous avons choisi de valider le programme en générant un jeu de données artificielles obéissant à un jeu de paramètres théoriques fixé, puis de lancer la procédure d'estimation sur ce jeu de données, pour enfin s'assurer que le jeu de paramètres estimé est raisonnablement proche du jeu de paramètres théorique. Cette opération supposant de savoir simuler des données obéissant à un LEYP donné, nous allons successivement présenter la distribution théorique conditionnelle des délais inter-événementiels et l'algorithme simulant

des dates d'événements s'enchaînant selon un LEYP. Nous présenterons enfin un exemple de résultat de validation de la procédure d'estimation.

5.3.1 Distribution conditionnelle du délai inter-événementiel

La variable aléatoire T_j étant définie comme l'instant où se produit la $j^{\text{ième}}$ défaillance, nous considérons le délai aléatoire $X_{j+1} = T_{j+1} - T_j$ séparant la $(j+1)^{\text{ième}}$ défaillance de la $j^{\text{ième}}$, conformément au schéma suivant :



La proposition suivante permet de caractériser la distribution conditionnelle de X_{j+1} :

Proposition 5.3. *La fonction de survie conditionnelle de X_{j+1} connaissant l'instant de la $j^{ième}$ défaillance est :*

$$\Pr\left\{X_{j+1} > x \mid T_j = t_j\right\} = \exp\left(-(1+\alpha j)\left[\Lambda(t_j + x) - \Lambda(t_j)\right]\right)$$

Preuve :

La survie conditionnelle cherchée est simplement la probabilité conditionnelle de ne pas avoir de défaillance dans $]t_j, t_j + x]$ sachant que $N(t_j) = j$, calculable en utilisant la proposition 4.1 :

$$\begin{aligned}
&\Pr\left\{X_{j+1} > x \mid T_j = t_j\right\} \\
&= \Pr\left\{N(t_j + x) - N(t_j) = 0 \mid N(t_j) = j\right\} \\
&= \exp\left(-(1+\alpha j)\left[\Lambda(t_j + x) - \Lambda(t_j)\right]\right)
\end{aligned}$$

□

5.3.2 Simulation numérique d'événements LEYP

Une séquence d'événements survenant selon un LEYP donné aux instants T_j, est facile à simuler sur ordinateur à partir d'une séquence pseudo-aléatoire de nombres uniformément distribués dans l'intervalle $[0, 1]$. Nous utilisons le procédé classique (voir par exemple [Ross, 1997]), dit de *la transformée inverse*, qui nécessite seulement de connaître l'expression analytique de l'inverse de la fonction de répartition des X_j.

Soit en effet une variable aléatoire U uniformément distribuée dans $[0, 1]$, et X une variable aléatoire réelle de fonction de répartition $F(x) = \Pr\{X \leqslant x\}$; F étant une fonction monotone

croissante, et puisque $U \sim \mathcal{U}_{[0,1]} \Rightarrow \Pr\{U \leqslant u\} = u$, $F^{-1}(U)$ suit alors la même distribution que X :$\Pr\left\{F^{-1}(U) \leqslant x\right\} = \Pr\left\{F(F^{-1}(U)) \leqslant F(x)\right\} = \Pr\{U \leqslant F(x)\} = F(x)$.

La proposition suivante permet de simuler par récurrence, en partant de $t_0 = 0$, des réalisations successives de T_j :

Proposition 5.4. *Si $(u_j : j \in \mathbb{N}^*)$ est une séquence d'aléas uniformes indépendants tirés dans $\mathcal{U}_{[0,1]}$, et t_j une réalisation de T_j, alors :*

$$t_{j+1} = \Lambda^{-1}\left(\Lambda(t_j) - \frac{\ln(1-u_{j+1})}{1+\alpha j}\right)$$

est une réalisation de T_{j+1}.

Preuve :

Appliquons la proposition 5.3 :

$$\begin{aligned} & 1 - \exp\left(-(1+\alpha j)\left[\Lambda(t_{j+1}) - \Lambda(t_j)\right]\right) = u_{j+1} \\ \Rightarrow & \Lambda(t_{j+1}) - \Lambda(t_j) = -\frac{\ln(1-u_{j+1})}{1+\alpha j} \\ \Rightarrow & t_{j+1} = \Lambda^{-1}\left(\Lambda(t_j) - \frac{\ln(1-u_{j+1})}{1+\alpha j}\right) \end{aligned}$$

□

5.4 Qualité d'ajustement du modèle LEYP

Nous n'avons pas encore tenté de développer formellement un test de qualité d'ajustement du modèle. Nous nous contentons pour l'instant de vérifier graphiquement que les courbes des taux de défaillance empirique et théorique sont raisonablement proches, et que les nombres de défaillances observés et prédits dans la fenêtre d'observation sont peu différents. La méthode de comparaison des taux de défaillance est détaillée à la sous-section 7.2.1.

5.5 Validation des prédictions du modèle LEYP

Un schéma original de validation a été mis au point par [Le Gat, 2002] dans le cadre du projet de recherche européen CareW du 5$^{\text{ième}}$ PCRD. Ce schéma est adapté aux études de modélisation du risque de casse de canalisations d'eau sous pression, où les canalisations sont le plus souvent observées depuis 1995, date moyenne à laquelle, en Europe de l'Ouest, les services gestionnaires des réseaux d'eau potable ont commencé à archiver leurs données de maintenance sur support électronique. La plupart des canalisations ont ainsi une fenêtre d'observation d'une

douzaine d'années, que l'on partage en une fenêtre dite « de calage » couvrant les 9 premières années, les 3 dernières années servant de fenêtre « de validation ». Les paramètres du LEYP sont ainsi estimés sur la fenêtre de calage ; les nombres moyens de casses sont ensuite prédits grâce au modèle sur la fenêtre de validation, où ils sont comparés aux nombres réellement observés. Comparer les prédictions aux réalisations observées revient à répondre aux questions :

- « les casses observées se concentrent-elles sur les canalisations ayant les nombres de casses prédits les plus élevés ? »
- « le nombre total de casses prédites pour un nombre statistiquement important de canalisations est-il proche du nombre total observé ? »

La première question est celle de la capacité du modèle à bien détecter les canalisations les plus à risque ; la seconde est celle du biais de prédiction.

5.5.1 Détection du risque

La méthode proposée par [Le Gat, 2002] repose sur l'étude du graphe de la proportion des casses évitables par la réhabilitation d'une proportion donnée du linéaire des canalisations les plus à risque au sens des prédictions du modèle. Nous supposons ici que n conduites sont observées, chacune dans l'intervalle $[a_i, c_i]$; les intervalles $[a_i, c_i]$ sont scindés aux fins de validation en un intervalle $[a_i, b_i]$ servant au calage du modèle, et en son complément $[b_i, c_i]$ servant à valider les prédictions. Les conduites sont supposées triées par valeurs décroissantes du taux de casse théorique, *i.e.* du nombre de casses prédit par le modèle LEYP rapporté à la longueur de la canalisation l_i. Nous notons :

- $k_j = N_j(c_j) - N_j(b_j)$ les nombres des casses observées,
- $\hat{k}_j = \mathrm{E}\left(N_j(c_j) - N_j(b_j) \mid N_j(b_j) - N_j(a_j)\right)$ les nombres des casses prédites,
- $r_i = \sum_{j=1}^{i} l_j / \sum_{j=1}^{n} l_j$ le rang relatif pondéré par la longueur de la $i^{\text{ième}}$ conduite la plus à risque,
- $f(r_i) = \sum_{j=1}^{i} k_j / \sum_{j=1}^{n} k_j$ la proportion de casses observées sur les i conduites les plus à risque,
- $\mathscr{A} = \sum_{i=1}^{n} l_i f(r_i) / \sum_{i=1}^{n} l_i$ l'aire sous la courbe de la fonction en escalier $f(r_i)$.

Plus l'aire $\mathscr{A}$ est proche de 1, *i.e.* plus la fonction en escalier $f(r_i)$ est proche du coin supérieur gauche, plus les casses observées dans les intervalles $[b_i, c_i]$ sont concentrées sur les canalisations dont le risque de casses au sens du modèle est le plus élevé. La valeur $f(r_i)$ estime en outre la proportion des casses totales qui peuvent être évitées en renouvelant la proportion r_i du linéaire totale la plus à risque au sens du modèle. C'est pourquoi le graphe de la fonction $f(r_i)$ est qualifié de *courbe de performance prédictive*. La figure 5.1 illustre la construction d'une telle courbe.

5.5.2 Diagnostic du biais de prédiction

Pour un groupe de canalisations suffisamment important, *e.g.* $i \geqslant 30$, l'absence de biais de prédiction peut être éprouvée en vérifiant que la somme des nombres des casses observées $\sum_{j=1}^{i} k_j$ se tient dans l'intervalle de confiance à 95% de la somme des nombres prédits $\sum_{j=1}^{i} \hat{k}_j$. Cet intervalle de confiance peut être estimé grâce à la normalité asymptotique de la somme de

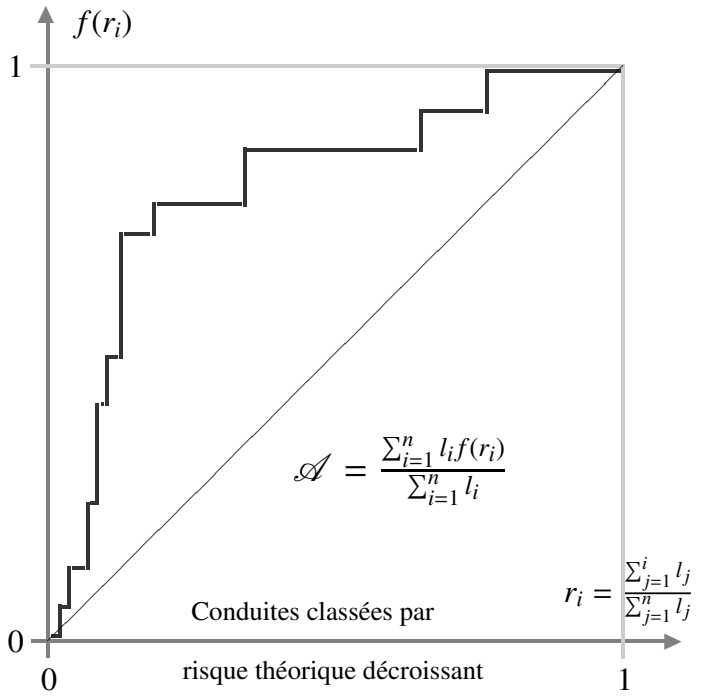


FIG. 5.1 – Courbe de performance prédictive

variables aléatoires indépendantes binomiales négatives de variance $\sum_{j=1}^{i} \mathrm{Var}\left(\hat{k}_j\right)$, à condition que les variances $\mathrm{Var}\left(\hat{k}_j\right)$ ne soient pas trop différentes les unes des autres.

Chapitre 6

Prise en compte des mises hors service de canalisations et biais de survie sélective

Bien que conduisant à un développement théorique de portée générale, le contexte des considérations à venir est celui du renouvellement des canalisations d'eau. Nous nous limiterons cependant à la considération de la mise hors service, le remplacement correspondant simplement à l'installation d'un nouvel objet, dont le processus de défaillances est modélisé indépendamment de celui de l'objet auquel il se substitue.

Considérons le fait que le système étudié a une durée de maintien en service limitée dans le temps. La durée de maintien en service est fréquemment dénommée « durée de vie » par les praticiens, dans un contexte technico-économique de recherche du temps optimal de remplacement ; notre opinion est cependant qu'il s'agit là d'un abus de langage, car une canalisation reste en service aussi longtemps que le gestionnaire du réseau décide de ne pas procéder à son remplacement, et il est donc peu pertinent de se référer implicitement à une notion de longévité intrinsèque. La durée de maintien en service est considérée comme une variable aléatoire ; sa distribution n'est cependant considérée dans notre étude que sur l'intervalle de temps qui précède la fenêtre d'observation. Au delà, les mises hors service de canalisations seront simulées de façon déterministe dans le cadre de comparaisons de stratégies de gestion patrimoniale, thème qui dépasse le cadre de notre travail.

La prise en compte avant la fenêtre d'observation d'une durée de maintien en service aléatoire a été motivée par l'exigence que le modèle LEYP reste utilisable en pratique avec des jeux de données comportant une proportion notable de canalisations très anciennes. L'étude de tels jeux de données montre en effet fréquemment que les individus les plus âgés manifestent singulièrement peu de défaillances. Cela est en fait dû à un phénomène de sélection : les canalisations posées anciennement, mais qui se sont illustrées par des défaillances répétées, ont le plus souvent été mises hors service pour cette raison, avant que ne débute la fenêtre d'observation. Les observations sont ainsi sujettes à un biais dit de survie sélective. De plus, au sein même de la fenêtre d'observation, la mise hors service sélective d'une canalisation à la suite d'une ou plusieurs défaillances censure à droite son observation d'une façon non indépendante du processus de défaillance.

6.1 Probabilité conditionnelle au maintien en service

Nous introduisons maintenant le modèle que nous appelons ζ-LEYP en ajoutant au modèle LEYP la fonction $\varsigma(t) \in [0,1]$ donnant la probabilité que suite à une défaillance à l'âge t la canalisation soit maintenue en service, *i.e.* fasse l'objet d'une simple réparation, la probabilité qu'elle soit mise hors service étant $1 - \varsigma(t)$.

Nous allons établir dans ce cadre la probabilité conditionnelle du nombre de défaillances dans un intervalle donné sachant que la canalisation n'a pas été mise hors service avant le début de cet intervalle, puis étendre le conditionnement à la connaissance du nombre de défaillances observées dans un intervalle antérieur.

Nous faisons d'abord l'hypothèse peu restrictive que $\varsigma(t)$ est une forme analytique constante par morceau. Nous sommes conduits à partitionner l'intervalle de temps allant de la pose de la canalisation à la fin de l'intervalle de prédiction en sous-intervalles de tailles quelconques que nous notons sous forme indicée $[a_{j-1}, a_j[, j \in \{1, 2, \ldots, n\}$; par convention $a_0 = 0$, et notre intervalle de prédiction antérieurement noté $[a, b[$ devient $[a_n, a_{n+1}[$. Nous avons par hypothèse :

$$\forall t \in [a_{j-1}, a_j[, \quad \exists \varsigma_j \in [0,1] : \quad \varsigma(t) = \varsigma_j$$

Pour étudier les propriétés probabilistes du modèle ζ-LEYP, il est commode, suivant le point de vue exposé par Cook et Lawless [Cook et Lawless, 1997], de considérer le modèle LEYP « complet » de fonction de comptage $N(t)$, comme aux chapitres précédents, sous condition que la mise hors service ne soit pas encore intervenue ; nous pouvons ainsi bénéficier pour nos démonstrations des résultats précédemment établis. Nous introduisons à cet effet la variable aléatoire T définie comme la durée de maintien en service (*i.e.* durée entre la pose de la canalisation et sa mise hors service).

Par souci de simplification d'écriture, nous notons respectivement $D_j = N(a_j-) - N(a_{j-1})$ le nombre de défaillances, et R_j l'indicatrice de l'occurrence de la mise hors service dans $[a_{j-1}, a_j[$. Nous utilisons aussi la notation allégée $\mu_n = \mu(a_n)$.

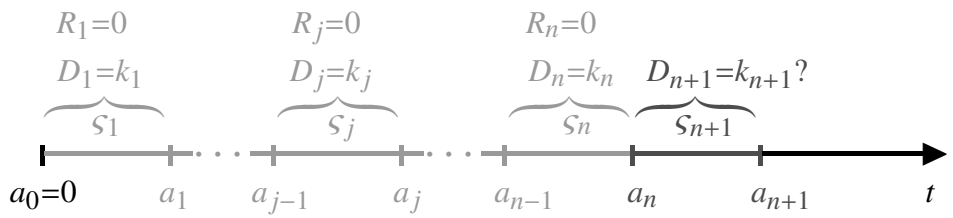
La distribution de la variable aléatoire T est liée aux R_j par la relation :

$$\Pr\{T > a_n\} = \Pr\left\{\sum_{j=1}^{n} R_j = 0\right\} \tag{6.1}$$

Nous avons du fait que la mise hors service est déterminée par l'occurrence des défaillances :

$$\Pr\left\{T > a_n \mid \left(D_j = k_j, j = 1, \ldots, n\right)\right\} = \prod_{j=1}^{n} \varsigma_j^{k_j} \tag{6.2}$$

Nous pouvons visualiser les notations adoptées dans le schéma suivant :



Nous cherchons la distribution conditionnelle de D_{n+1} sachant $T > a_n$. Comme nous utiliserons le théorème de Bayes pour calculer $\Pr\{D_{n+1} = k_{n+1} \mid T > a_n\}$, commençons par calculer $\Pr\{T > a_n \mid D_{n+1} = k_{n+1}\}$:

Proposition 6.1.

$$\Pr\{T > a_n \mid D_{n+1} = k_{n+1}\} = \left(\frac{\mu_{n+1} - \mu_n + 1}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{\alpha^{-1}+k_{n+1}}$$

Preuve :

Montrons que la relation est vraie pour $n = 1$:

$$\Pr\{T > a_1 \mid D_2 = k_2\} = \sum_{k_1=0}^{\infty} \Pr\{T > a_1 \mid D_1 = k_1, D_2 = k_2\} \Pr\{D_1 = k_1 \mid D_2 = k_2\}$$

Or, de par la détermination de la mise hors service et en utilisant (6.2) :

$$\Pr\{T > a_1 \mid D_1 = k_1, D_2 = k_2\} = \Pr\{T > a_1 \mid D_1 = k_1\} = \varsigma_1^{k_1}$$

En outre selon la proposition 4.6 :

$$\Pr\{D_1 = k_1 \mid D_2 = k_2\} = \frac{\Gamma(\alpha^{-1} + k_2 + k_1)}{\Gamma(\alpha^{-1} + k_2)k_1!} \left(\frac{\mu_2 - \mu_1 + 1}{\mu_2}\right)^{\alpha^{-1}+k_2} \left(\frac{\mu_1 - 1}{\mu_2}\right)^{k_1}$$

D'où, en utilisant (4.2) :

$$\begin{aligned}
&\Pr\{T > a_1 \mid D_2 = k_2\} \\
&= \left(\frac{\mu_2 - \mu_1 + 1}{\mu_2}\right)^{\alpha^{-1}+k_2} \sum_{k_1=0}^{\infty} \frac{\Gamma(\alpha^{-1} + k_2 + k_1)}{\Gamma(\alpha^{-1} + k_2)k_1!} \varsigma_1^{k_1} \left(\frac{\mu_1 - 1}{\mu_2}\right)^{k_1} \\
&= \left(\frac{\mu_2 - \mu_1 + 1}{\mu_2}\right)^{\alpha^{-1}+k_2} \left(1 - \frac{\varsigma_1(\mu_1 - 1)}{\mu_2}\right)^{-(\alpha^{-1}+k_2)} \\
&= \left(\frac{\mu_2 - \mu_1 + 1}{\mu_2 - \varsigma_1(\mu_1 - 1)}\right)^{\alpha^{-1}+k_2}
\end{aligned}$$

Faisons maintenant l'hypothèse que la proposition est vraie à l'ordre $n-1$:

$$\Pr\{T > a_{n-1} \mid D_n = k_n\} = \left(\frac{\mu_n - \mu_{n-1} + 1}{\mu_n - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{\alpha^{-1}+k_n}$$

Nous pouvons écrire à l'ordre n :

$$\begin{aligned}
&\Pr\{T > a_n \mid D_{n+1} = k_{n+1}\} \\
&= \Pr\{T > a_{n-1}, R_n = 0 \mid D_{n+1} = k_{n+1}\} \\
&= \sum_{k_n=0}^{\infty} \Pr\{T > a_{n-1}, R_n = 0 \mid D_n = k_n, D_{n+1} = k_{n+1}\} \Pr\{D_n = k_n \mid D_{n+1} = k_{n+1}\} \\
&= \sum_{k_n=0}^{\infty} \Pr\{R_n = 0 \mid T > a_{n-1}, D_n = k_n, D_{n+1} = k_{n+1}\} \Pr\{T > a_{n-1} \mid D_n = k_n, D_{n+1} = k_{n+1}\} \times \\
&\qquad \times \Pr\{D_n = k_n \mid D_{n+1} = k_{n+1}\}
\end{aligned}$$

Appliquons l'hypothèse de récurrence :

$$\begin{aligned}
&\Pr\{T > a_{n-1} \mid D_n = k_n, D_{n+1} = k_{n+1}\} = \Pr\{T > a_{n-1} \mid D_n + D_{n+1} = k_n + k_{n+1}\} \\
&= \left(\frac{\mu_{n+1} - \mu_{n-1} + 1}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{\alpha^{-1}+k_n+k_{n+1}}
\end{aligned}$$

En outre :

$$\Pr\{R_n = 0 \mid T > a_{n-1}, D_n = k_n, D_{n+1} = k_{n+1}\} = \Pr\{R_n = 0 \mid D_n = k_n\} = \varsigma_n^{k_n}$$

De plus, en appliquant la proposition 4.6 :

$$\Pr\{D_n = k_n \mid D_{n+1} = k_{n+1}\} = \frac{\Gamma(\alpha^{-1} + k_n + k_{n+1})}{\Gamma(\alpha^{-1} + k_{n+1})\, k_n!} \left(\frac{\mu_{n+1} - \mu_n + 1}{\mu_{n+1} - \mu_{n-1} + 1}\right)^{\alpha^{-1}+k_{n+1}} \left(\frac{\mu_n - \mu_{n-1}}{\mu_{n+1} - \mu_{n-1} + 1}\right)^{k_n}$$

D'où :

$$\begin{aligned}
&\Pr\{T > a_n \mid D_{n+1} = k_{n+1}\} \\
&= \left(\frac{\mu_{n+1} - \mu_n + 1}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{\alpha^{-1}+k_{n+1}} \sum_{k_n=0}^{\infty} \frac{\Gamma(\alpha^{-1} + k_n + k_{n+1})}{\Gamma(\alpha^{-1} + k_{n+1})\, k_n!} \left(\frac{\varsigma_n(\mu_n - \mu_{n-1})}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{k_n}
\end{aligned}$$

Utilisons à nouveau (4.2) :

$$= \left(\frac{\mu_{n+1} - \mu_n + 1}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{\alpha^{-1}+k_{n+1}} \left(1 - \frac{\varsigma_n(\mu_n - \mu_{n-1})}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{-(\alpha^{-1}+k_{n+1})}$$

$$= \left(\frac{\mu_{n+1} - \mu_n + 1}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1} + k_{n+1}} \qquad \square$$

Nous pouvons maintenant établir la :

Proposition 6.2. *En supposant que* $\varsigma_j = 1$ *pour* $j \geqslant n+1$ *:*

$$[D_{n+1} = k_{n+1} \mid T > a_n] \sim \mathcal{NB}\left(\alpha^{-1}, \frac{\mu_n - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})} \right)$$

Preuve :

Appliquons le théorème de Bayes :

$$\Pr\{D_{n+1} = k_{n+1} \mid T > a_n\} = \frac{\Pr\{T > a_n \mid D_{n+1} = k_{n+1}\} \Pr\{D_{n+1} = k_{n+1}\}}{\sum_{k_{n+1}=0}^{\infty} \Pr\{T > a_n \mid D_{n+1} = k_{n+1}\} \Pr\{D_{n+1} = k_{n+1}\}}$$

D'après la proposition 6.1 :

$$\Pr\{T > a_n \mid D_{n+1} = k_{n+1}\} = \left(\frac{\mu_{n+1} - \mu_n + 1}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1} + k_{n+1}}$$

Et d'après la proposition 4.3 :

$$\Pr\{D_{n+1} = k_{n+1}\} = \frac{\Gamma(\alpha^{-1} + k_{n+1})}{\Gamma(\alpha^{-1})\, k_{n+1}!} \left(\frac{1}{\mu_{n+1} - \mu_n + 1} \right)^{\alpha^{-1}} \left(\frac{\mu_{n+1} - \mu_n}{\mu_{n+1} - \mu_n + 1} \right)^{k_{n+1}}$$

D'où :

$$\begin{aligned} &\Pr\{T > a_n \mid D_{n+1} = k_{n+1}\} \Pr\{D_{n+1} = k_{n+1}\} \\ &= \left(\frac{1}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1}} \frac{\Gamma(\alpha^{-1} + k_{n+1})}{\Gamma(\alpha^{-1})\, k_{n+1}!} \left(\frac{\mu_{n+1} - \mu_n}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{k_{n+1}} \end{aligned}$$

Et donc, en utilisant (4.2) :

$$\begin{aligned} &\sum_{k_{n+1}=0}^{\infty} \Pr\{T > a_n \mid D_{n+1} = k_{n+1}\} \Pr\{D_{n+1} = k_{n+1}\} \\ &= \left(\frac{1}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1}} \left(1 - \frac{\mu_{n+1} - \mu_n}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{-\alpha^{-1}} \\ &= \left(\frac{1}{\mu_n - \sum_{j=1}^{n} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1}} \end{aligned}$$

D'où finalement :

$$\begin{aligned}&\Pr\{D_{n+1} = k_{n+1} \mid T > a_n\}\\ &= \frac{\Gamma(\alpha^{-1} + k_{n+1})}{\Gamma(\alpha^{-1})\, k_{n+1}!} \left(\frac{\mu_n - \sum_{j=1}^{n} \varsigma_j(\mu_j - \mu_{j-1})}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{\alpha^{-1}} \left(\frac{\mu_{n+1} - \mu_n}{\mu_{n+1} - \sum_{j=1}^{n} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{k_{n+1}}\end{aligned}$$ □

Etendons la proposition 6.2 en conditionnant aussi sur le nombre de défaillances observées dans l'intervalle précédent :

Proposition 6.3. *En supposant que $\varsigma_j = 1$ pour $j \geqslant n+1$:*

$$[D_{n+1} \mid D_n = k_n, T > a_n] \sim \mathcal{NB}\left(\alpha^{-1} + k_n, \frac{\mu_n - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}\right)$$

Preuve :

Appliquons le théorème de Bayes :

$$\begin{aligned}&\Pr\{D_{n+1} = k_{n+1} \mid D_n = k_n, T > a_n\}\\ &= \frac{\Pr\{T > a_n \mid D_n = k_n, D_{n+1} = k_{n+1}\} \Pr\{D_{n+1} = k_{n+1} \mid D_n = k_n\}}{\sum_{k_{n+1}=0}^{\infty} \Pr\{T > a_n \mid D_n = k_n, D_{n+1} = k_{n+1}\} \Pr\{D_{n+1} = k_{n+1} \mid D_n = k_n\}}\end{aligned}$$

Or d'après la proposition 6.1 :

$$\begin{aligned}&\Pr\{T > a_n \mid D_n = k_n, D_{n+1} = k_{n+1}\}\\ &= \Pr\{T > a_{n-1}, R_n = 0 \mid D_n = k_n, D_{n+1} = k_{n+1}\}\\ &= \Pr\{R_n = 0 \mid T > a_{n-1}, D_n = k_n, D_{n+1} = k_{n+1}\} \Pr\{T > a_{n-1} \mid D_n = k_n, D_{n+1} = k_{n+1}\}\\ &= \varsigma_n^{k_n} \left(\frac{\mu_{n+1} - \mu_{n-1} + 1}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j(\mu_j - \mu_{j-1})}\right)^{\alpha^{-1} + k_n + k_{n+1}}\end{aligned}$$

Et d'après la proposition 4.5 :

$$\begin{aligned}&\Pr\{D_{n+1} = k_{n+1} \mid D_n = k_n\}\\ &= \frac{\Gamma(\alpha^{-1} + k_n + k_{n+1})}{\Gamma(\alpha^{-1} + k_n)\, k_{n+1}!} \left(\frac{\mu_n - \mu_{n-1} + 1}{\mu_{n+1} - \mu_{n-1} + 1}\right)^{\alpha^{-1} + k_n} \left(\frac{\mu_{n+1} - \mu_n}{\mu_{n+1} - \mu_{n-1} + 1}\right)^{k_{n+1}}\end{aligned}$$

Ainsi, en utilisant (4.2) :

$$\sum_{k_{n+1}=0}^{\infty} \Pr\{T > a_n \mid D_n = k_n, D_{n+1} = k_{n+1}\} \Pr\{D_{n+1} = k_{n+1} \mid D_n = k_n\}$$

$$
\begin{aligned}
&= \varsigma_n^{k_n} \left(\frac{\mu_n - \mu_{n-1} + 1}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1}+k_n} \sum_{k_{n+1}=0}^{\infty} \frac{\Gamma(\alpha^{-1} + k_n + k_{n+1})}{\Gamma(\alpha^{-1} + k_n)\, k_{n+1}!} \left(\frac{\mu_{n+1} - \mu_n}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{k_{n+1}} \\
&= \varsigma_n^{k_n} \left(\frac{\mu_n - \mu_{n-1} + 1}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1}+k_n} \left(1 - \frac{\mu_{n+1} - \mu_n}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{-(\alpha^{-1}+k_n)} \\
&= \varsigma_n^{k_n} \left(\frac{\mu_n - \mu_{n-1} + 1}{\mu_n - \sum_{j=1}^{n-1} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1}+k_n}
\end{aligned}
$$

Et donc :

$$
\begin{aligned}
&\Pr\{D_{n+1} = k_{n+1} \mid D_n = k_n, T > a_n\} \\
&= \frac{\Gamma(\alpha^{-1} + k_n + k_{n+1})}{\Gamma(\alpha^{-1} + k_n)\, k_{n+1}!} \left(\frac{\mu_n - \sum_{j=1}^{n-1} \varsigma_j (\mu_j - \mu_{j-1})}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{\alpha^{-1}+k_n} \left(\frac{\mu_{n+1} - \mu_n}{\mu_{n+1} - \sum_{j=1}^{n-1} \varsigma_j (\mu_j - \mu_{j-1})} \right)^{k_{n+1}}
\end{aligned}
$$

□

Remarque 6.1. *La proposition 6.3 a ceci de remarquable que la distribution conditionnelle de* $[D_{n+1} \mid D_n = k_n, T > a_n]$ *ne dépend pas de* ς_n *; la censure des observations dans* $[a_{n-1}, a_n]$ *par le processus de défaillance lui-même peut donc être considérée comme indépendante.* △

La proposition 6.4 suivante généralise la proposition 6.3 en considérant une fonction $\varsigma(t)$ continue, et en faisant tendre vers l'infini le nombre des intervalles $[a_{j-1}, a_j[$ tout en faisant tendre leur longueur vers 0 :

Proposition 6.4.

$$
[N(c) - N(b+) \mid N(b) - N(a) = m, T > b] \sim \mathcal{NB}\left(\alpha^{-1} + m, \frac{\mu(b) - \int_0^a \varsigma(t) \mathrm{d}\mu(t)}{\mu(c) - \int_0^a \varsigma(t) \mathrm{d}\mu(t)} \right)
$$

Preuve :

Réécrivons la proposition 6.3 en renommant a_{n-1} en a, a_n en b, a_{n+1} en c, et partitionnons $[0, a[$ en n sous-intervalles :

$$
[0, a[= \bigcup_{j=0}^{n-1} \left[\frac{ja}{n}, \frac{(j+1)a}{n} \right[
$$

Nous avons alors :

$$
\begin{aligned}
&[N(c) - N(b+) \mid N(b) - N(a) = m, T > b] \\
&\sim \mathcal{NB}\left(\alpha^{-1} + m, \frac{\mu(b) - \sum_{j=0}^{n-1} \varsigma\left(\frac{ja}{n}\right) \left(\mu\left(\frac{(j+1)a}{n}\right) - \mu\left(\frac{ja}{n}\right) \right)}{\mu(c) - \sum_{j=0}^{n-1} \varsigma\left(\frac{ja}{n}\right) \left(\mu\left(\frac{(j+1)a}{n}\right) - \mu\left(\frac{ja}{n}\right) \right)} \right)
\end{aligned}
$$

Or :

$$\lim_{n\to+\infty}\sum_{j=0}^{n-1}\varsigma\left(\frac{ja}{n}\right)\left(\mu\left(\frac{(j+1)a}{n}\right)-\mu\left(\frac{ja}{n}\right)\right)=\int_0^a\varsigma(t)\mathrm{d}\mu(t)$$

□

6.2 Prédiction conditionnelle

En matière de prévision, nous considérerons des probabilités conditionnelles au maintien en service jusqu'à la fin de l'intervalle d'observation $[a,b]$, et supposerons, comme à la proposition 6.4, que la probabilité de mise hors service suite à une défaillance postérieure à b est nulle. Rappelons en effet que notre but est de simuler des remplacements, afin de comparer des stratégies de gestion patrimoniale, et non d'extrapoler dans le futur les pratiques passées.

Le calcul de prédiction connaissant le nombre de défaillances dans la fenêtre d'observation et sachant qu'aucune de ces défaillances n'a conduit à la mise hors service de la conduite nécessite d'établir la :

Proposition 6.5. *Soit un système obéissant à un ζ-LEYP défini par* (6.1)*, en supposant que $\varsigma(t)=1$ pour $t\geqslant b$. La distribution conditionnelle de $N(d)-N(c)$, sachant que $N(b)-N(a)=m$ et $T>b=0$, avec $0<a<b<c<d$, est binomiale négative :*

$$\begin{aligned}&[N(d)-N(c)\mid N(b)-N(a)=m,T>b=0]\\&\sim\mathcal{NB}\left(\alpha^{-1}+m,\frac{\mu(b)-\int_0^a\varsigma(u)\mathrm{d}\mu(u)}{\mu(d)-\mu(c)+\mu(b)-\int_0^a\varsigma(u)\mathrm{d}\mu(u)}\right)\end{aligned}$$

Preuve :

Appliquons la formule des probabilités totales :

$$\begin{aligned}&\Pr\{N(d)-N(c)=k\mid N(b)-N(a)=m,T>b=0\}\\&=\sum_{j=0}^{\infty}\Pr\{N(d)-N(c)=k\mid N(c-)-N(b+)=j,N(b)-N(a)=m,T>b=0\}\times\\&\qquad\times\Pr\{N(c-)-N(b+)=j\mid N(b)-N(a)=m,T>b=0\}\end{aligned}$$

Suivant l'hypothèse selon laquelle $\varsigma(t)=1$ pour $t\geqslant b$, et appliquant la proposition 6.4 :

$$\begin{aligned}&\Pr\{N(d)-N(c)=k\mid N(c-)-N(b+)=j,N(b)-N(a)=m,T>b=0\}\\&=\Pr\{N(d)-N(c)=k\mid N(c-)-N(a)=m,T>c=0\}\end{aligned}$$

$$= \frac{\Gamma(\alpha^{-1}+m+j+k)}{\Gamma(\alpha^{-1}+m+j)\,k!}\left(\frac{\mu(c)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}{\mu(d)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}+m+j}\left(\frac{\mu(d)-\mu(c)}{\mu(d)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{k}$$

De plus, toujours selon la proposition 6.4 :

$$\begin{aligned}&\Pr\{N(c-)-N(b+)=k \mid N(b)-N(a)=m, T>b=0\}\\ &= \frac{\Gamma(\alpha^{-1}+m+j)}{\Gamma(\alpha^{-1}+m)\,j!}\left(\frac{\mu(b)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}{\mu(c)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}+m}\left(\frac{\mu(c)-\mu(b)}{\mu(c)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{j}\end{aligned}$$

D'où, en utilisant l'équation (4.2) :

$$\begin{aligned}&\Pr\{N(d)-N(c)=k \mid N(b)-N(a)=m, T>b=0\}\\ &= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+m)\,k!}\left(\frac{\mu(b)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}{\mu(d)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}+m}\left(\frac{\mu(d)-\mu(c)}{\mu(d)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{k}\times\\ &\quad\times\sum_{j=0}^{\infty}\frac{\Gamma(\alpha^{-1}+m+k+j)}{\Gamma(\alpha^{-1}+m+k)\,j!}\left(\frac{\mu(c)-\mu(b)}{\mu(d)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{j}\\ &= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+m)\,k!}\left(\frac{\mu(b)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}{\mu(d)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}+m}\left(\frac{\mu(d)-\mu(c)}{\mu(d)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{k}\times\\ &\quad\times\left(1-\frac{\mu(c)-\mu(b)}{\mu(d)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{-(\alpha^{-1}+m+k)}\\ &= \frac{\Gamma(\alpha^{-1}+m+k)}{\Gamma(\alpha^{-1}+m)\,k!}\left(\frac{\mu(b)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}{\mu(d)-\mu(c)+\mu(b)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}+m}\times\\ &\quad\times\left(\frac{\mu(d)-\mu(c)}{\mu(d)-\mu(c)+\mu(b)-\int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{k}\end{aligned}$$

□

6.3 Probabilité de maintien en service

Depuis le milieu des années 90, la prise de conscience des gestionnaires de réseaux et des maîtres d'ouvrages de la nécessité d'une bonne gestion patrimoniale de ces infrastructures s'est accompagnée d'un débat opposant les tenants d'une stratégie de renouvellement des canalisations assurant le maintien d'un âge moyen pour la population des conduites, aux tenants d'une stratégie de renouvellement *a minima*, n'assurant le remplacement que des seules canalisations manifestant des défaillances répétées. Il nous paraît clair qu'on ne peut pas définir *a priori* un âge limite applicable à une catégorie de conduites, même homogène en matériau et diamètre,

sans tenir compte des conditions dans lesquelles elles ont été, et seront, exploitées. Il semble en outre légitime de garder en service des canalisations même très anciennes, dès lors qu'elles ne sont responsables ni d'interruptions de service à répétition, ni d'une dégradation de la qualité de l'eau qu'elles transportent. La durée de maintien en service d'une canalisation résulte *in fine* d'un choix du gestionnaire qui, à l'extrême, peut préférer conserver une conduite peu fiable au détriment de la qualité du service qu'elle assure, ou au contraire remplacer « par sécurité » une conduite saine mais considérée comme trop âgée. A cet égard, le modèle ζ-LEYP peut nous permettre d'estimer, au moins pour les catégories de conduites les plus anciennes, les fonctions de survie résultant de la stratégie de renouvellement passée modélisée par la fonction $\varsigma(t)$. Il nous suffit pour cela d'établir la :

Proposition 6.6.

$$\Pr\{T > a\} = \left(\mu(a) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)^{-\alpha^{-1}}$$

Preuve :

Considérons les instants 0<a<b, et appliquons la proposition 6.1 :

$$\Pr\{T > a \mid N(b) - N(a) = k\} = \left(\frac{\mu(b) - \mu(a) + 1}{\mu(b) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}+k}$$

Appliquons la formule des probabilités totales :

$$\begin{aligned}
\Pr\{T > a\} &= \sum_{k=0}^{\infty} \Pr\{T > a \mid N(b) - N(a) = k\} \Pr\{N(b) - N(a) = k\} \\
&= \sum_{k=0}^{\infty} \left(\frac{\mu(b) - \mu(a) + 1}{\mu(b) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}+k} \frac{\Gamma(\alpha^{-1} + k)}{\Gamma(\alpha^{-1})k!} \left(\frac{1}{\mu(b) - \mu(a) + 1}\right)^{\alpha^{-1}} \left(\frac{\mu(b) - \mu(a)}{\mu(b) - \mu(a) + 1}\right)^{k} \\
&= \left(\frac{1}{\mu(b) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}} \sum_{k=0}^{\infty} \frac{\Gamma(\alpha^{-1} + k)}{\Gamma(\alpha^{-1})k!} \left(\frac{\mu(b) - \mu(a)}{\mu(b) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{k} \\
&= \left(\frac{1}{\mu(b) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}} \left(1 - \frac{\mu(b) - \mu(a)}{\mu(b) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)}\right)^{-\alpha^{-1}} \\
&= \left(\mu(a) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)^{-\alpha^{-1}}
\end{aligned}$$

□

6.4 Forme analytique de $\varsigma(t)$

Nous nous trouvons maintenant confrontés au choix d'une forme analytique pour $\varsigma(t)$. La forme en double exponentielle :

$$\varsigma(t) = \exp(-\exp(\zeta_0 + \zeta_1 t)), \zeta_0 \in \mathbb{R}, \zeta_1 \in \mathbb{R}_+ \tag{6.3}$$

a été retenue ; elle rend simplement compte de la tendance « naturelle » des gestionnaires de réseaux à plutôt réparer les canalisations jeunes, et plutôt remplacer les canalisations anciennes suite à leur défaillance. Le cas $\zeta_1 = 0$ peut en outre être utile pour corriger le biais de survie sélective dans un échantillon de canalisations suffisamment défaillantes pour que ce biais soit sensible, mais trop jeunes pour que ζ_1 puisse être estimé correctement.

Comme nous le verrons au chapitre 8 la forme (6.3) permet de modéliser convenablement le risque de défaillance des fontes grises, dont la période de pose couvre les années 1850 à 1970, et dont la complexité du comportement est illustrée par la figure 8.1. La figure 6.1 porte les courbes théoriques du taux de défaillance et de la probabilité de maintien en service en fonction de l'âge obtenues avec des valeurs de paramètres typiques de celles observées pour la fonte grise étudiée au chapitre 8 ; le cas particulier retenu pour cette figure est celui d'un tronçon de 100 m, de diamètre 100 mm, posé après 1945 et installé sous chaussée ; les paramètres du modèles ζ-LEYP sont $\alpha = 1.9$, $\delta = 1.2$, $\zeta_0 = -3.1$, $\zeta_1 = 0.011$ (an^{-1}), $\beta_0 = -7.6$, $\beta_1 = 0.55$, $\beta_2 = -0.0017$ (mm^{-1}), $\beta_3 = -0.75$ (Pose 1850-1889), $\beta_4 = -0.52$ (Pose 1890-1930), $\beta_5 = -0.67$ (Pose 1931-1945) et $\beta_6 = 0.08$ (Sous chaussée).

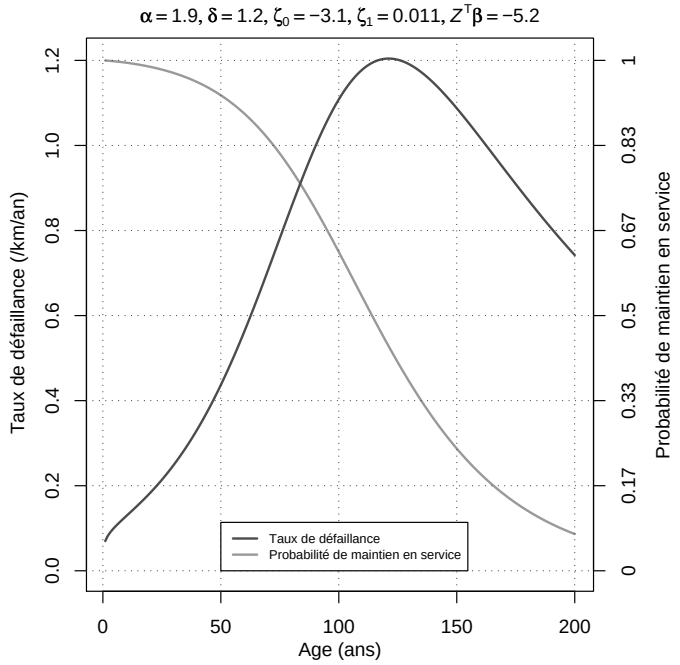


FIG. 6.1 – Taux de défaillance et probabilité de maintien en service théoriques avec biais de survie sélective

La courbe théorique du taux de défaillance en fonction de l'âge résulte de la :

Proposition 6.7. *L'intensité non conditionnelle du ζ-LEYP est :*

$$\mathrm{E}\left(\mathrm{d}N_{\cdot}(t) \mid N_2(t-) = 0\right) = \frac{\lambda(t)\mu(t)\mathrm{d}t}{\mu(t) - \int_0^t \varsigma(u)\mathrm{d}\mu(u)}$$

Preuve :

$$\begin{aligned}
&\mathrm{E}\left(\mathrm{d}N_{\cdot}(t) \mid N_2(t-) = 0\right) \\
&= \Pr\left\{N_{\cdot}(t+\mathrm{d}t) - N_{\cdot}(t) = 1 \mid N_2(t-) = 0\right\}
\end{aligned}$$

soit en appliquant la proposition 6.4 avec $a = b = t, c = t + \mathrm{d}t, m = 0$:

$$\begin{aligned}
&= \frac{\Gamma(\alpha^{-1}+1)}{\Gamma(\alpha^{-1})\,1!}\left(\frac{\mu(t) - \int_0^t \varsigma(u)\mathrm{d}\mu(u)}{\mu(t+\mathrm{d}t) - \int_0^t \varsigma(u)\mathrm{d}\mu(u)}\right)^{\alpha^{-1}}\left(\frac{\mu(t+\mathrm{d}t) - \mu(t)}{\mu(t+\mathrm{d}t) - \int_0^t \varsigma(u)\mathrm{d}\mu(u)}\right) \\
&= \frac{\alpha^{-1}\mathrm{d}\mu(t)}{\mu(t) - \int_0^t \varsigma(u)\mathrm{d}\mu(u)} \\
&= \frac{\lambda(t)\mu(t)\mathrm{d}t}{\mu(t) - \int_0^t \varsigma(u)\mathrm{d}\mu(u)}
\end{aligned}$$

□

La probabilité théorique de maintien en service est donnée par la proposition 6.6 de la section 6.3.

L'effet remarquable de la survie sélective est que le taux de défaillance n'est plus monotone croissant, mais peut décroître au delà d'un certain âge.

6.5 Vraisemblance du ζ-LEYP

Pour construire la fonction de vraisemblance des paramètres d'un processus de comptage avec censure à droite dépendante, nous allons considérer un processus bidimensionnel avec deux types de défaillances :

– type 1 - défaillance suivie d'une réparation,
– type 2 - défaillance suivie de la mise hors service de la canalisation.

Notons respectivement $N_1(t)$ et $N_2(t)$ les processus de comptage des deux types de défaillances, et leur somme $N_{\cdot}(t) = N_1(t) + N_2(t)$. Nous considérons qu'à chaque défaillance la probabilité que celle-ci soit de type 1 est $\varsigma(t)$, et $1 - \varsigma(t)$ de type 2. Appelons ζ-LEYP le processus LEYP bidimensionnel spécifié par la :

Définition 6.1. *Le modèle ζ-LEYP est défini par le système d'équations :*

$$\begin{aligned}
&\forall t \in \mathbb{R}_+, \alpha \in \mathbb{R}_+^*, \varsigma(t) \in [0,1], \lambda(t) \in \mathbb{R}_+^* \\
&\begin{cases}
N_1(0) = 0, N_2(0) = 0 \\
\mathrm{E}\left(\mathrm{d}N_1(t) \mid N_1(t-), N_2(t-)\right) = \varsigma(t)(1 - N_2(t-))(1 + \alpha N_1(t-))\lambda(t)\mathrm{d}t \\
\mathrm{E}\left(\mathrm{d}N_2(t) \mid N_1(t-), N_2(t-)\right) = (1 - \varsigma(t))(1 - N_2(t-))(1 + \alpha N_1(t-))\lambda(t)\mathrm{d}t
\end{cases}
\end{aligned}$$

Ce modèle rend compte de l'observation, qui par nature impose une intensité de défaillances nulle après la disparition de la canalisation.

Afin de construire la fonction de vraisemblance du vecteur $\boldsymbol{\theta}$ du modèle ζ-LEYP connaissant une séquence de défaillances observées aux instants t_j, $j \in \{1, \ldots, m\}$ dans l'intervalle $[a, b]$ (en l'absence de défaillance, $m = 0$), nous avons besoin de calculer l'intensité conditionnelle du processus $N.()$ à tout instant $t \in [a, b]$ connaissant le nombre de défaillances observées sur $[a, t[$ et sachant que la défaillance la plus récente n'a pas entraîné la mise hors service de la canalisation. Cette dernière condition s'écrit $N_2(t-) = 0$ et assure en pratique que la canalisation est encore observable juste avant t.

Considérant une fonction $\varsigma(t)$ continue, montrons la proposition suivante relative à la fonction d'intensité du ζ-LEYP :

Proposition 6.8. *L'intensité conditionnelle d'un ζ-LEYP à t, connaissant ses défaillances sur $[a, t[$ et sachant qu'il est observable juste avant t, est :*

$$\forall t \in [a, b], \forall k \in \{0, \ldots, m\},$$

$$\text{(i)} \quad \mathrm{E}\left(\mathrm{d}N.(t) \mid N.(t-) - N.(a) = k, N_2(t-) = 0\right) = \left(\alpha^{-1} + k\right) \mathrm{d} \ln \left(\mu(t) - \int_0^a \varsigma(u) \mathrm{d}\mu(u)\right)$$

$$\begin{aligned} \text{(ii)} \quad & \mathrm{E}\left(\mathrm{d}N_1(t) \mid N.(t-) - N.(a) = k, N_2(t-) = 0\right) \\ & = \varsigma(t) \mathrm{E}\left(\mathrm{d}N.(t) \mid N.(t-) - N.(a) = k, N_2(t-) = 0\right) \end{aligned}$$

$$\begin{aligned} \text{(iii)} \quad & \mathrm{E}\left(\mathrm{d}N_2(t) \mid N.(t-) - N.(a) = k, N_2(t-) = 0\right) \\ & = (1 - \varsigma(t)) \mathrm{E}\left(\mathrm{d}N.(t) \mid N.(t-) - N.(a) = k, N_2(t-) = 0\right) \end{aligned}$$

Preuve :

Concernant le point (i) :

$$\begin{aligned} & \mathrm{E}\left(\mathrm{d}N.(t) \mid N.(t-) - N.(a) = k, N_2(t-) = 0\right) \\ & = \Pr\left\{N.(t + \mathrm{d}t) - N.(t) = 1 \mid N.(t-) - N.(a) = k, N_2(t-) = 0\right\} \end{aligned}$$

soit en appliquant la proposition 6.4 avec $b = t, c = t + \mathrm{d}t$:

$$\begin{aligned} & = \frac{\Gamma(\alpha^{-1} + k + 1)}{\Gamma(\alpha^{-1} + k)\, 1!} \left(\frac{\mu(t) - \int_0^a \varsigma(u) \mathrm{d}\mu(u)}{\mu(t + \mathrm{d}t) - \int_0^a \varsigma(u) \mathrm{d}\mu(u)}\right)^{\alpha^{-1}+k} \left(\frac{\mu(t + \mathrm{d}t) - \mu(t)}{\mu(t + \mathrm{d}t) - \int_0^a \varsigma(u) \mathrm{d}\mu(u)}\right) \\ & = \left(\alpha^{-1} + k\right) \frac{\mathrm{d}\mu(t)}{\mu(t) - \int_0^a \varsigma(u) \mathrm{d}\mu(u)} \\ & = \left(\alpha^{-1} + k\right) \mathrm{d} \ln \left(\mu(t) - \int_0^a \varsigma(u) \mathrm{d}\mu(u)\right) \end{aligned}$$

Les points (ii) et (iii) sont évidents de par la définition 6.1. □

En suivant l'exposé de [Andersen *et al.*, 1993], la vraisemblance du vecteur de paramètres $\boldsymbol{\theta}$ d'un ζ-LEYP connaissant ses défaillances dans $[a, b]$ est donnée par la :

Définition 6.2. *la vraisemblance $L(\boldsymbol{\theta})$ du vecteur de paramètres $\boldsymbol{\theta}$ d'un ζ-LEYP, connaissant la séquence de défaillances $((t_j, N_2(t_j)) \in [a,b], j \in \{1,\dots,m\})$ est :*

$$L(\boldsymbol{\theta}) = \prod_{t\in[a,b]} \left(1 - \mathrm{E}\left(\mathrm{d}N_{\cdot}(t) \mid N_{\cdot}(t-) - N_{\cdot}(a), N_2(t-) = 0\right)\right)^{1-\Delta N_{\cdot}(t)} \times$$
$$\times \prod_{h=1}^{2} \mathrm{E}\left(\mathrm{d}N_h(t) \mid N_{\cdot}(t-) - N_{\cdot}(a), N_2(t-) = 0\right)^{\Delta N_h(t)}$$
$$\text{où :} \quad \Delta N_h(t) = N_h(t) - N_h(t-)$$

La vraisemblance $L(\boldsymbol{\theta})$ peut être calculée grâce à la proposition :

Proposition 6.9. *la vraisemblance $L(\boldsymbol{\theta})$ du vecteur de paramètres $\boldsymbol{\theta}$ d'un ζ-LEYP, connaissant la séquence de défaillances $((t_j, N_2(t_j)) \in [a,b], j \in \{1,\dots,m\})$ se calcule comme :*

$$L(\boldsymbol{\theta}) = \alpha^m \frac{\Gamma(\alpha^{-1}+m)}{\Gamma(\alpha^{-1})} \frac{\left(\mu(a) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)^{\alpha^{-1}}}{\left(\mu(b) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)^{\alpha^{-1}+m}} \prod_{j=1}^{m} \lambda(t_j)\mu(t_j)\varsigma(t_j)^{1-N_2(t_j)}(1-\varsigma(t_j))^{N_2(t_j)}$$

Preuve :

Calculons d'abord grâce à la proposition 6.8 le produit des intensités aux instants t_j :

$$\begin{aligned}
&\prod_{t\in[a,b]} \prod_{h=1}^{2} \mathrm{E}\left(\mathrm{d}N_h(t) \mid N_{\cdot}(t-) - N_{\cdot}(a), N_2(t-) = 0\right)^{\Delta N_h(t)} \\
&= \prod_{j=1}^{m} \prod_{h=1}^{2} \mathrm{E}\left(\mathrm{d}N_h(t) \mid N_{\cdot}(t-) - N_{\cdot}(a), N_2(t-) = 0\right)^{\Delta N_h(t)} \\
&= \prod_{j=1}^{m} \frac{(\alpha^{-1} + (j-1))\alpha\lambda(t_j)\mu(t_j)\varsigma(t_j)^{1-N_2(t_j)}(1-\varsigma(t_j))^{N_2(t_j)}}{\mu(t_j) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)} \\
&= \alpha^m \frac{\Gamma(\alpha^{-1}+m)}{\Gamma(\alpha^{-1})} \frac{\prod_{j=1}^{m} \lambda(t_j)\mu(t_j)\varsigma(t_j)^{1-N_2(t_j)}(1-\varsigma(t_j))^{N_2(t_j)}}{\prod_{j=1}^{m}\left(\mu(t_j) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)}
\end{aligned}$$

Calculons ensuite le produit des probabilités de ne pas avoir de défaillance entre les t_j :

$$\begin{aligned}
&\prod_{t\in[a,b]} \left(1 - \mathrm{E}\left(\mathrm{d}N_{\cdot}(t) \mid N_{\cdot}(t-) - N_{\cdot}(a), N_2(t-) = 0\right)\right)^{1-\Delta N_{\cdot}(t)} \\
&= \prod_{j=0}^{m} \prod_{t\in]t_j,t_{j+1}[} \left(1 - \mathrm{E}\left(\mathrm{d}N_{\cdot}(t) \mid N_{\cdot}(t-) - N_{\cdot}(a) = j, N_2(t-) = 0\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \prod_{j=0}^{m} \exp\Big(- \int_{t_j}^{t_{j+1}} \mathrm{E}\left(\mathrm{d}N_{\cdot}(t) \mid N_{\cdot}(t-) - N_{\cdot}(a) = j, N_2(t-) = 0\right) \Big) \\
&= \prod_{j=0}^{m} \exp\Big(- \int_{t_j}^{t_{j+1}} (\alpha^{-1} + j)\mathrm{d}\ln\left(\mu(t) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)\Big) \\
&= \prod_{j=0}^{m} \left(\frac{\mu(t_j) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)}{\mu(t_{j+1}) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)} \right)^{\alpha^{-1}+j} \\
&= \frac{\left(\mu(a) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)^{\alpha^{-1}}}{\left(\mu(b) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)^{\alpha^{-1}+m}} \prod_{j=1}^{m} \left(\mu(t_j) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)
\end{aligned}
$$

Et simplifions finalement par $\prod_{j=1}^{m} \left(\mu(t_j) - \int_0^a \varsigma(u)\mathrm{d}\mu(u)\right)$. □

6.6 Estimation des paramètres d'un ζ-LEYP

L'estimation des paramètres d'un ζ-LEYP en maximisant le logarithme de la vraisemblance exprimée par la proposition 6.9 reste numériquement praticable par la méthode de Levenberg-Marquardt évoquée en section 5.2. Cependant, le terme $\int_0^a \varsigma(u)\mathrm{d}\mu(u)$ peut ne pas être explicitement calculable et doit donc être approché par une quadrature, telle que celle de Gauss-Legendre (voir [Abramowitz et Stegun, 1972]). Cela rend difficile le calcul explicite du vecteur gradient et de la matrice hessienne ; ces quantités doivent donc être estimées numériquement, par exemple par la méthode des *forward differences* expliquée dans le chapitre relatif à la procédure de SAS/STAT NLMIXED dans [SAS Institute Inc., 1999]. L'inconvénient de la mise en oeuvre de ces approximations est la forte augmentation du temps de calcul, rançon du gain de souplesse indéniable qu'elles introduisent dans la définition du modèle.

6.7 Etude graphique de la log-vraisemblance du ζ-LEYP

Dans cette section, nous vérifions de façon graphique que la convexité de la fonction de log-vraisemblance du ζ-LEYP est bien adaptée à la recherche de son maximum par une méthode numérique de type quasi-Newton. Nous avons à cet effet généré un jeu de données aléatoires, de description de canalisations et de dates de défaillance, pour 20 000 conduites posées entre 1900 et 2005, et observées entre le 1er janvier 1990 et le 31 décembre 2006. Par souci de simplification, nous avons considéré le cas d'une covariable unique de type indicatrice, appelée Z_1 dans la suite, prenant la valeur 1 dans un cas sur deux. Chaque défaillance simulée conduit aléatoirement à une simple réparation de la conduite ou à sa mise hors service. Les paramètres choisis pour le ζ-LEYP générant les défaillances et les mises hors service sont :

$$\alpha = +2.5$$

$$\delta = +1.3$$
$$\zeta_0 = -3.0$$
$$\zeta_1 = +3.0$$
$$\beta_0 = -0.5$$
$$\beta_1 = +0.3$$

Ces valeurs sont « réalistes ». Elles sont par ailleurs adaptées à des durées exprimées en siècles, unité de temps qui permet d'éviter les problèmes numériques dans les calculs impliquant la fonction $\mu(t) = e^{\alpha\Lambda(t)}$ (*cf.* équations (3.2) et (4.1)) avec l'intensité de Yule-Weibull-Cox définie par 3.2, qui peut provoquer des débordements de réels en double précision de forme $\exp(t^\delta)$.

6.7.1 Simulation de données

Le schéma porté à la figure 6.2 représente l'algorithme de simulation du jeu de données. Les données sont enregistrées dans deux fichiers : l'un pour la description des canalisations, l'autre pour les dates de défaillance. Le fichier des canalisations porte un enregistrement par conduite avec les champs suivants :

– identifiant de la conduite ;
– date de pose ;
– date de mise hors service ;
– valeur de Z_1.

Le fichier des défaillances porte un enregistrement par défaillance avec les champs suivants :

– identifiant de la conduite ;
– date de défaillance.

Le programme informatique a été écrit en Java (Java™2 Platform Standard Edition 5.0) ; il y est fait usage de 4 séquences pseudo-aléatoires de distribution uniforme, désignées ci-après u_j, $j = 1, \ldots, 4$.

Pour chaque canalisation, le processus est initialisé par le choix de l'année de pose DP comme un entier aléatoire u_1 uniformément distribué entre 1900 et 2005 (bornes incluses), et par le choix de la valeur de la covariable Z_1 comme un entier aléatoire u_2 équiprobablement distribué dans $\{0, 1\}$. Le processus se poursuit en générant des dates de défaillance successives D, fonctions des réels u_3 uniformément distribués dans $[0, 1]$, selon la méthode exposée en sous-section 5.3.2. Chaque date de défaillance comprise entre les dates de début d'observation DD et de fin d'observation DF se traduit par un enregistrement dans le fichier des défaillances. Le processus stoppe si l'une des deux conditions suivantes est rencontrée :

– la date de défaillance simulée est postérieure à la date d'arrêt des observations DF ;
– la défaillance entraîne la mise hors service de la conduite, *i.e.* si le réel u_4 uniformément distribué dans $[0, 1]$ satisfait la condition $u_4 > \varsigma(D - DP)$ (*cf.* définition 6.1).

Dans le premier cas, un enregistrement est écrit dans le fichier descriptif des conduites, avec une date de mise hors service laissée vide.

Dans le second cas :

– si la défaillance intervient avant le début de la fenêtre d'observation, la mise hors service suite à cette défaillance n'est pas observée, et aucun enregistrement concernant cette

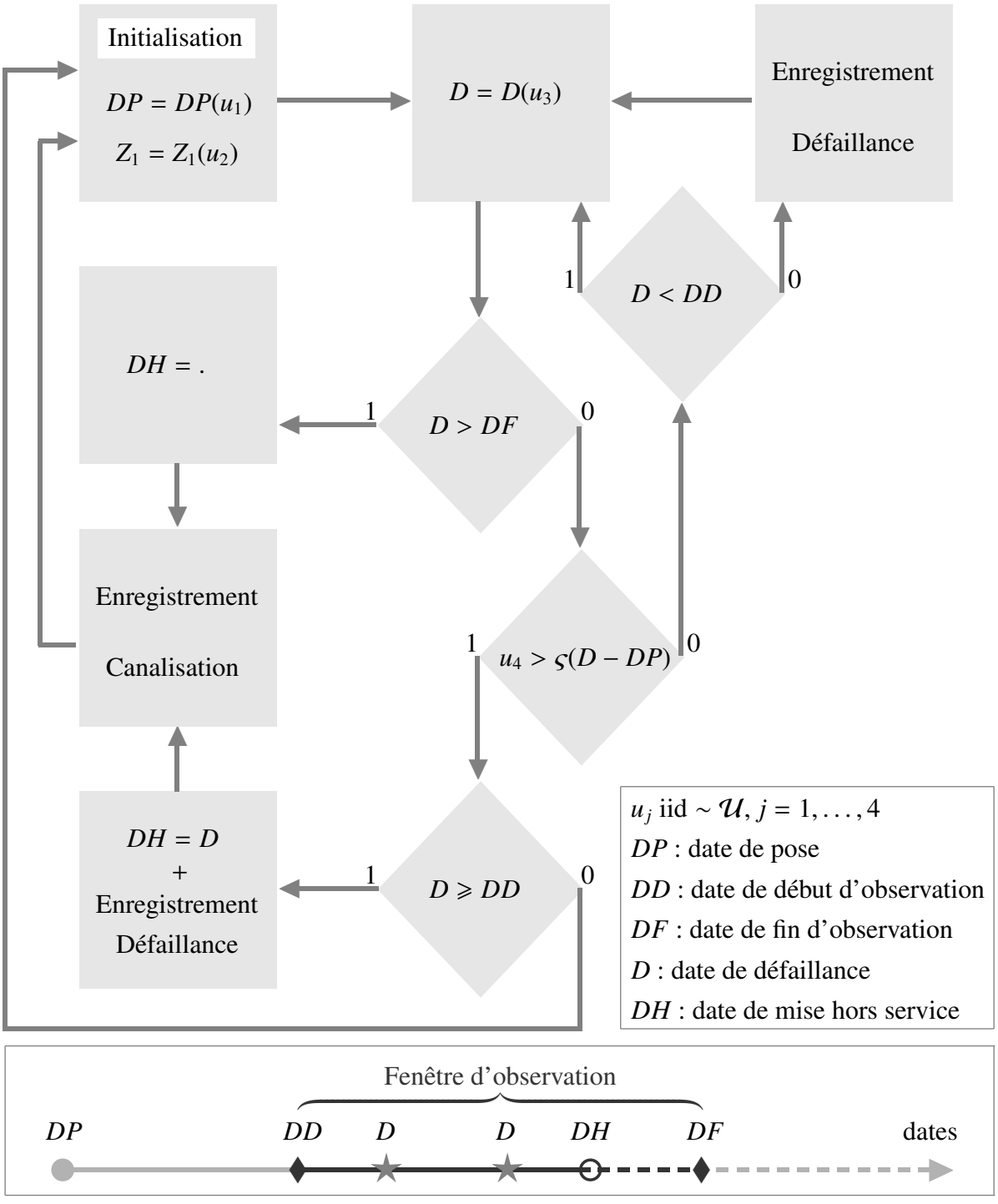


FIG. 6.2 – Organigramme de la simulation de données selon un ζ-LEYP

conduite n'est créé ;
– si la défaillance intervient dans la fenêtre d'observation, elle est enregistrée dans le fichier des défaillances, et un enregistrement est créé dans le fichier des canalisations avec une date de mise hors service égale à celle de la défaillance.

6.7.2 Résultats de calage du ζ-LEYP

Le jeu de données résultant concerne 18 815 conduites, 1 185 conduites sur les 20 000 considérées initialement ayant été mises hors service avant le début de la fenêtre d'observation. Les défaillances observées sont au nombre de 3 648, et concernent 2 872 tronçons. La distribution du nombre de défaillances observées par tronçon est portée au tableau 6.1.

Nb défaillances par tronçon	Nombre de tronçons	Proportion (%)
0	15 943	84.74
1	2 293	12.19
2	431	2.29
3	109	0.58
4	29	0.16
5	10	0.05
Total tronçons	18 815	100.00
Total défaillances	3 648	

TAB. 6.1 – Distribution des nombres de défaillances par tronçon

Les valeurs obtenues par maximisation de la vraisemblance pour les estimations des paramètres d'un ζ-LEYP sur ce jeu de données artificielles sont portées au tableau 6.2. Il est à noter que les valeurs estimées sont très proches des valeurs théoriques choisies pour les simulations ; ces dernières sont d'ailleurs toujours incluses dans les intervalles de confiance à 95 % des valeurs estimées.

Remarque 6.2. *Les simulations numériques ont été répétées un grand nombre de fois, avec des jeux de paramètres théoriques variés, que la maximisation de la vraisemblance permet toujours de retrouver, avec un écart aléatoire minime. Cela autorise une certaine confiance dans l'utilisation de la procédure calculatoire d'estimation des paramètres du ζ-LEYP.* ◁

6.7.3 Etude graphique de la fonction de vraisemblance

L'allure de la fonction de vraisemblance est étudiée graphiquement à partir du jeu de données artificielles, en traçant, sous système R, pour chacun des paramètres le graphe de la fonction dans un intervalle de valeurs du paramètre concerné incluant la valeur optimale issue du calage, les autres paramètres restant fixés à leur valeur optimale. Ces graphes sont portés à la figure 6.3. La valeur maximale prise par la log-vraisemblance sur ce jeu de données artificielles

Paramètre	Valeur théorique	Valeur estimée	IC 95 % de la valeur estimée
α	+2.5	+2.390	[+2.145, +2.662]
δ	+1.3	+1.334	[+1.284, +1.392]
ζ_0	-3.0	-3.001	[−3.222, −2.780]
ζ_1	+3.0	+2.996	[+2.715, +3.306]
β_0	-0.5	-0.477	[−0.531, −0.422]
β_1	+0.3	+0.299	[+0.245, +0.352]

TAB. 6.2 – Valeurs théoriques et estimées des paramètres du ζ-LEYP étudié par simulations numériques

est −3 831.42. Les 6 graphes ont une forme nettement convexe. Le graphe afférent au paramètre δ, responsable de l'accroissement de l'intensité de défaillance avec l'âge de la canalisation, montre un maximum sensiblement moins accusé que celui des autres paramètres, suggérant la possibilité de difficultés d'ajustement lorsque l'effet du vieillissement est peu marqué, *i.e.* lorsque δ est proche de 1 ; la suite du présent travail devrait ainsi inclure une étude de sensibilité relative à l'écart à la valeur 1 qu'il est possible de mettre en évidence pour ce paramètre, en fonction du nombre de défaillances observées dans le jeu de données.

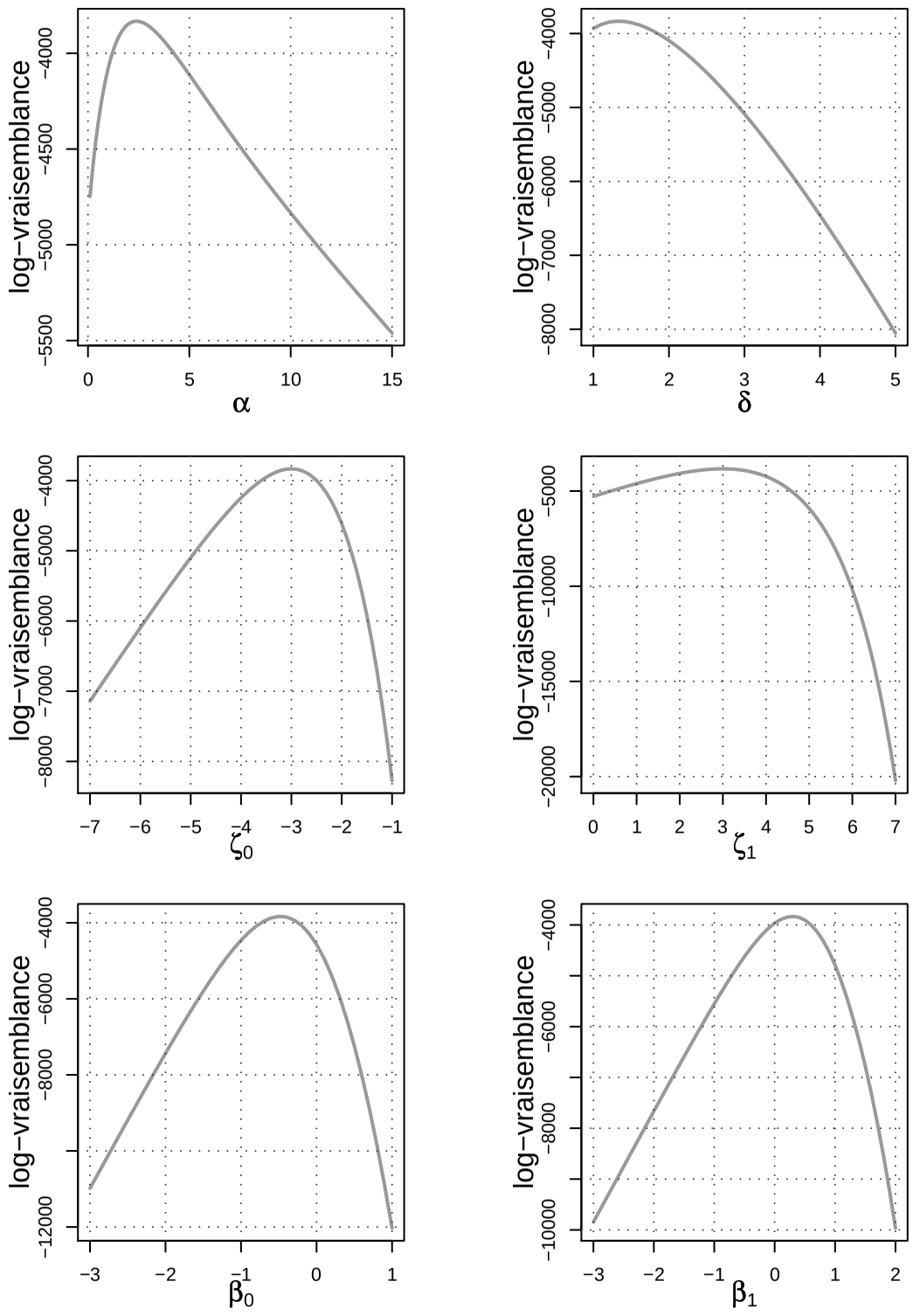


FIG. 6.3 – Graphes de la log-vraisemblance du ζ-LEYP autour des valeurs optimales des paramètres

Deuxième partie

Applications

Chapitre 7

Introduction

Nous nous proposons d'illustrer l'intérêt de l'utilisation du modèle LEYP, pour analyser des données de maintenance de canalisations en vue d'établir des priorités de renouvellement, en nous appuyant sur les données du réseau de distribution d'eau potable d'une grosse collectivité urbaine du nord de la France. Ce site d'étude permet d'éprouver les performances prédictives du modèle LEYP sur de gros échantillons de canalisations, présentant une large variété de caractéristiques (matériaux, diamètres, âges, *etc.*). Nous avons choisi de stratifier l'étude des casses par matériau, en considérant *a priori* que chaque type de matériau réagit de façon propre aux agressions physico-chimiques de son environnement, et que cela doit se traduire par des plages de valeurs spécifiques pour les paramètres d'un ζ-LEYP. Les cinq principaux types de matériaux représentés dans cette étude sont :

- la fonte grise,
- la fonte ductile,
- l'acier,
- le béton âme-tôle,
- le polyéthylène haute densité (PEHD).

La figure 7.1 porte la distribution des linéaires pour ces matériaux. Nous avons choisi de centrer nos applications numériques sur les deux sortes de fontes, qui représentent près de 80 % du linéaire du réseau.

La chronique de défaillances disponible couvre exhaustivement les années 1995 à 2006. Nous avons choisi d'utiliser les neufs premières années 1995 à 2003 de cette série pour caler chacun des modèles (par matériau), puis de calculer les nombres attendus de défaillances pour chaque conduite et de comparer ces prédictions aux nombres de défaillances réellement observés sur les trois dernières années 2004 à 2006, en utilisant la méthode de validation décrite en section 5.5.

7.1 Les tronçons

L'individu statistique considéré est le *tronçon*, *i.e.* l'ensemble des tuyaux unitaires, assemblés par des joints, homogènes en matériau, diamètre et date de pose, et formant un segment de réseau généralement délimité à ses deux extrémités par des organes de sectionnement (vannes). Le tronçon est le composant élémentaire du réseau de canalisations ; sa description est dispo-

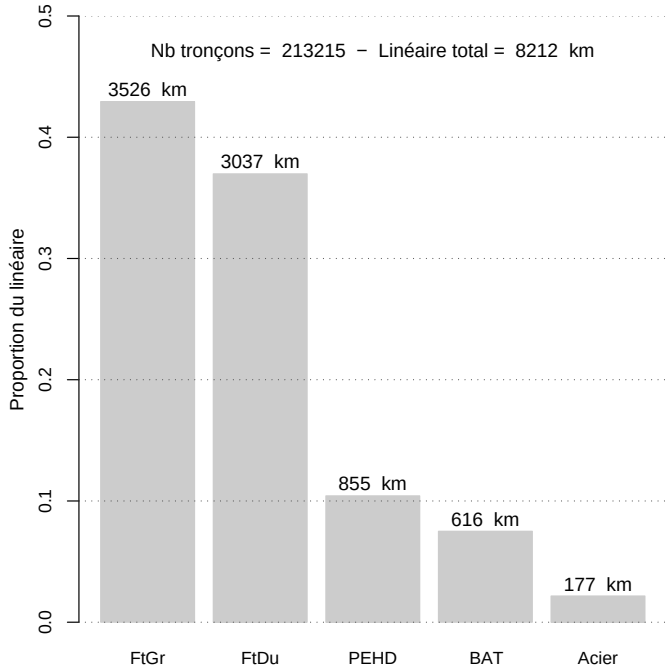


FIG. 7.1 – Distribution des linéaires par matériau

nible dans une base de données du Système d'Information Géographique (SIG), tenue à jour par le service gestionnaire du système de distribution d'eau potable. Parmi les éléments descriptifs du tronçon, certains sont considérés *a priori* comme étant potentiellement explicatifs du niveau de risque de défaillance ; la pertinence de l'utilisation de ces facteurs de risque comme covariables du modèle est éprouvée par la méthode explicitée à la section 5.2.

7.2 Les défaillances

En accord avec les responsables techniques chargés de la gestion patrimoniale du réseau d'eau étudié, sont indifféremment considérées comme *défaillances* :

– les fuites manifestes en surface,
– et les fuites non manifestes réparées suite à une campagne de recherche active,

qu'elles surviennent :

– sur corps de conduite,
– ou sur joint.

La répartition détaillée des défaillances par type est donnée pour chaque matériau étudié au chapitre le concernant.

L'indicateur de fiabilité d'un linéaire de canalisation le plus utilisé en pratique est le taux de défaillance. Ce dernier est le plus souvent abusivement dénommé *taux de casses*, mais comme nous le verrons aux chapitres 8 et 9, la rupture du corps de la conduite n'est pas le seul type de défaillance possible. Cet indicateur est défini comme le nombre de défaillances survenant dans un intervalle de temps *suffisamment court* rapporté à la longueur de cet intervalle et au linéaire des canalisations concernées, et généralement exprimé en nombre de défaillances par km et par

an. La question de l'influence du vieillissement sur le taux de défaillances étant centrale dans notre travail, nous illustrons plus loin par des graphes par matériau la relation entre le taux de défaillances et l'âge des canalisations. La méthode utilisée pour estimer empiriquement le taux de défaillances fait l'objet de la sous-section 7.2.1.

7.2.1 L'estimation du taux de défaillance empirique

L'estimateur empirique classiquement utilisé pour étudier l'adéquation d'un processus de comptage théorique « modèle » aux chroniques de défaillances observées sur une collection d'individus est celui de Nelson-Aalen, exposé dans [Andersen *et al.*, 1993], que nous redéfinissons cependant ici comme le nombre cumulé $\hat{\Psi}(t)$ de défaillances par unité de longueur de tronçon en fonction de l'âge, et à le comparer à $\mathrm{E}\Psi(t)$, la quantité équivalente estimée par le modèle. Il convient ici de souligner que nous différons de la pratique classique en ce que les nombres de défaillances sont rapportés au linéaire de canalisations exposées au risque, plutôt qu'au nombre d'individus à risque ; cela ne change rien aux propriétés statistiques de l'estimateur, car le linéaire exposé au risque ne diffère du nombre d'individus à risque que par un facteur d'échelle non aléatoire.

Considérons un n-échantillon de canalisations, de longueurs l_i, dont les fonctions de comptage sont notées $N_i(t)$, observées sur les intervalles d'âges $[a_i, b_i]$. Chaque tronçon est observé défaillir m_i fois (m_i peut être nul) aux instants $t_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i$, réalisations des variables aléatoires T_{ij}. Si tous les a_i étaient nuls, nous aurions :

$$\forall t \in [0, \max_{i=1}^{n}(b_i)] :$$
$$\hat{\Psi}(t) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathrm{I}\left(t_{ij} \leq t\right) \left(\sum_{i=1}^{n} \mathrm{I}\left(b_i \geq t\right) l_i \right)^{-1}$$

et

$$\mathrm{E}\Psi(t) = \sum_{i=1}^{n} \mathrm{E}N_i(t) \mathrm{I}\left(a_i \leq t \leq b_i\right) \left(\sum_{i=1}^{n} \mathrm{I}\left(a_i \leq t \leq b_i\right) l_i \right)^{-1}$$

Du fait que a_i n'est nul que pour une petite minorité de tronçons, posés après le début de la fenêtre d'observation du réseau, la seule quantité directement accessible est la différence $N_i(t) - N_i(a_i)$, qui ne dépend donc pas seulement de l'âge du tronçon mais aussi de l'âge en début de fenêtre d'observation, variable selon les tronçons ; $\hat{\Psi}(t)$ ne peut donc pas être calculé.

Il est cependant possible de contrôler l'adéquation du modèle sur l'estimation du taux de défaillances empirique en fonction de l'âge des canalisations, calculée en lissant les incréments de l'estimateur de Nelson-Aalen $\hat{\Psi}(t)$ par un noyau d'Epanechnikov, suivant en cela les recommandations de [Andersen *et al.*, 1993]. Le lissage permet de produire une estimation annuelle du taux de défaillance malgré la rareté des défaillances à cette échelle de temps. La largeur de bande Δt choisie pour la fonction noyau est de 4 ans. La fonction noyau d'Epanechnikov est définie par :

$$K(x) = 0.75(1 - x^2)\mathrm{I}\left(\mid x \mid \leq 1\right), x \in \mathbb{R}$$

Nous calculons donc les incréments lissés $\hat{\psi}(t)$ de $\hat{\Psi}(t)$ comme :

$$\hat{\psi}(t) = \left(\sum_{i=1}^{n} \sum_{j=1}^{m_i} K\left(\frac{t - t_{ij}}{\Delta t} \right) \right) \left(\sum_{i=1}^{n} \mathrm{I}\left(a_i \leq t \leq b_i\right) l_i \right)^{-1} (\Delta t)^{-1}$$

ainsi que leur variance :

$$\mathrm{Var}\left(\hat{\psi}(t)\right) = \left(\sum_{i=1}^{n} \sum_{j=1}^{m_i} K^2\left(\frac{t - t_{ij}}{\Delta t} \right) \right) \left(\sum_{i=1}^{n} \mathrm{I}\left(a_i \leq t \leq b_i\right) l_i \right)^{-2} (\Delta t)^{-2}$$

L'adéquation du modèle aux observations de défaillances pourra être controlée en comparant graphiquement $\hat{\psi}(t)$, assorti d'un intervalle de confiance basé sur $\mathrm{Var}\left(\hat{\psi}(t)\right)$, à $\mathrm{E}\psi(t)$ calculé comme :

$$\mathrm{E}\psi(t) = \sum_{i=1}^{n} \mathrm{E}\mathrm{d}N_i(t) \mathrm{I}\left(a_i \leq t \leq b_i\right) \left(\sum_{i=1}^{n} \mathrm{I}\left(a_i \leq t \leq b_i\right) l_i \right)^{-1}$$

La comparaison peut aussi être effectuée par rapport à l'espérance conditionnelle de $\psi(t)$:

$$\mathrm{E}\left(\psi(t) \mid N(t-)\right) = \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{d}N_i(t) \mid N_i(t-)\right) \mathrm{I}\left(a_i \leq t \leq b_i\right) \left(\sum_{i=1}^{n} \mathrm{I}\left(a_i \leq t \leq b_i\right) l_i \right)^{-1}$$

7.2.2 Le taux de défaillance annuel moyen observé

Le taux moyen annuel de défaillance est porté, pour tous les matériaux confondus, au tableau 7.1. Ce taux moyen est ici simplement calculé comme le nombre total de défaillances observées une année donnée rapporté au linéaire de réseau en service cette année là. Le climat est suspecté d'être largement responsable de la forte variation inter-annuelle observée (du simple au double).

7.3 Les canalisations mises hors service

Chaque année une petite fraction du linéaire du réseau est mise hors service, et le plus souvent remplacée par des canalisations neuves. Deux causes principales sont traditionnellement distinguées comme déterminant la mise hors service d'une conduite :

– les travaux de voirie,
– les défaillances à répétition de la conduite.

Les travaux de voirie peuvent :

– nécessiter le déplacement de la conduite (par exemple, en cas d'installation d'un tramway),
– ou inciter, voire contraindre, le gestionnaire du réseau d'eau à s'assurer qu'aucune intervention sur la conduite ne sera nécessaire à moyen terme.

Les défaillances à répétition entraînent :

– une dégradation marquée de la qualité de service pour les usagers dont l'approvisionnement en eau dépend de la conduite considérée,
– des travaux de réparation occasionnant à leur tour leur lot de nuisances aux riverains ou au trafic.

Contrairement à la mise hors service pour cause de défaillances répétées, la mise hors service pour cause de travaux de voirie pourrait être *a priori* considérée comme non informative au regard du processus de défaillances. Il convient cependant de remarquer que les défaillances répétées d'une conduite et les travaux qu'elles génèrent se traduisent immanquablement par une détérioration de la chaussée, et incitent *in fine* les services de la voirie à décider la rénovation de la voie. Par ailleurs, confronté à une décision de rénovation prise par le service de la voirie, le gestionnaire du service d'eau doit souvent décider s'il en profite ou non pour remplacer la conduite qui dessert la voie concernée ; l'occurrence d'une défaillance dans un passé proche (dans les cinq dernières années) incite généralement au remplacement. La distinction entre les deux causes de mise hors service d'une conduite est donc loin d'être évidente, et la mise hors service doit souvent être considérée comme informative, justifiant ainsi de toujours tenter en première approche de caler un ζ-LEYP plutôt qu'un LEYP simple, quitte à revenir dans un second temps sur cette hypothèse, si les paramètres ζ s'avèrent être non significatifs.

Les mises hors service annuelles de conduites, en moyenne de l'ordre de 0.35 % du linéaire, sont portées, pour tous les matériaux confondus, au tableau 7.1. Ce tableau laisse aussi apparaître une croissance marquée du réseau (une trentaine de km par an) du fait de la densification et de l'expansion du tissu urbain.

Remarque 7.1. *Dans la base de données utilisée, seule l'année de mise hors service du tronçon est reportée, la date précise n'étant pas archivée au jour près. Les mises hors service de canalisations, principalement aux fins de renouvellement à l'identique et plus rarement de redimensionnement, font l'objet de travaux décidés annuellement et programmés indépendamment de la réparation des défaillances. Le cadre théorique du modèle ζ-LEYP, qui considère qu'une défaillance entraîne soit une réparation soit une mise hors service, est une simplification de la réalité qui conduit à substituer à la date de mise hors service la date de la défaillance la plus récente, ou à défaut de défaillance dans la fenêtre d'observation, le 31 décembre de l'année de mise hors service portée dans la base de données.* △

7.4 Les covariables disponibles

Nous pouvons distinguer deux grandes catégories de covariables :
- celles dites « générales »,
- et celles dites « locales ».

Les covariables générales sont définies dans quasiment tous les service d'eau de la même façon et leur disponibilité dans la base de données « tronçons » du service gestionnaire est presque toujours assurée. Celles considérées dans notre application sont :
- le type de matériau,
- la longueur,
- le diamètre,
- la profondeur d'installation,
- la cote altimétrique.

Contrairement aux covariables générales, la définition des covariables locales et leur disponibilité reflètent largement la culture technique du service gestionnaire ; ce sont pour cette étude :
- la période de pose,

- le type de joints entre tuyaux élémentaires,
- le type d'occupation du sol au dessus de la canalisation,
- le type d'encaissant.

Remarque 7.2. *L'utilisation de covariables locales compromet la transposabilité du modèle d'un site à l'autre, y compris au sein d'une même aire géographique. Cela n'enlève cependant rien à son intérêt pour une utilisation locale. Il est par ailleurs loin d'être évident qu'un modèle n'utilisant que des covariables générales soit « brutalement » transposable d'un service où il a été calé, à un autre.* ◁

Remarque 7.3. *Il est à noter que nous ne disposons pas dans cette étude de covariables décrivant la pression de service de la canalisation, ni le trafic dans la voie sous laquelle elle est installée.* ◁

7.4.1 La longueur du tronçon

La longueur du tronçon est éminemment variable, de quelques décimètres, à plusieurs centaines de mètres, avec une distribution proche de la Loi de Pareto (la densité de probabilité a une allure dite « en i »), qui varie en outre selon le matériau. S'il est « naturel » de penser que le nombre de défaillances par unité de temps doit augmenter avec la longueur du tronçon, il est cependant loin d'être évident que la relation soit proportionnelle. Nous verrons plus loin que le taux de défaillance est plutôt proportionnel à une puissance de la longueur, variable selon le matériau, tournant toujours grossièrement autour de 0.5. Il est donc important de noter que l'expression quasi-universellement utilisée en pratique de la fiabilité d'une canalisation, à savoir le « taux de défaillance » exprimé en nombre de défaillances par unités de temps et de longueur, *n'est pas indépendante de la longueur de la canalisation*.

7.4.2 Le diamètre des tuyaux

Il s'agit pour la plupart des matériaux du diamètre intérieur du tuyau, à l'exception notable du PEHD où il est d'usage d'afficher le diamètre extérieur. Le diamètre est une caractéristique quantitative de la canalisation qui varie de façon discrète. Si on observe le plus souvent que le taux de défaillance tend à baisser lorsque le diamètre croît, la relation n'est pas forcément proportionnelle, ni complètement monotone. Il semble à l'usage pertinent de grouper les diamètres voisins en classes consistantes, et d'utiliser comme covariables les indicatrices de ces classes. La distribution du diamètre est très variable selon le matériau.

7.4.3 La profondeur d'installation

Bien que la profondeur techniquement conseillée se situe entre 0.80 et 1.20 m, cette caractéristique apparaît comme très variable sur le service étudié. Un nombre non négligeable de canalisations se trouvent en outre à la surface du sol ou au dessus, installées sur berceaux dans des ouvrages d'art (ponts, galeries). Il ne paraît raisonnable d'envisager un effet de ce facteur sur le taux de défaillance que pour les canalisations enterrées ; il semble alors pertinent de partitionner l'intervalle de variation de la profondeur, et d'utiliser comme covariables les indicatrices des classes obtenues.

7.4.4 La cote altimétrique

La zone géographique desservie par le réseau d'eau étudié se situe dans une plage d'altitudes allant de 105 m à 245 m au dessus du niveau de la mer. Comme en témoigne la figure 7.2, l'altitude 120 m est largement la mieux représentée. On peut s'attendre *a priori* à ce que l'altitude influe sur le régime de pression de service des canalisations, et donc indirectement sur leur risque de défaillance ; il est toutefois difficile de préjuger de la forme du lien éventuel, le réseau comportant plusieurs paliers de pression. Là encore, nous procéderons en partitionnant l'intervalle de variation de l'altitude, puis en utilisant comme covariables les indicatrices des classes obtenues. La topographie en plaines et coteaux parfois pentus fait que la distribution de l'altitude est multimodale et déséquilibrée, induisant des regroupements en classes de largeurs hétérogènes. L'altitude n'étant pas renseignée dans le SIG pour 283 tronçons, soit environ 12 km de canalisations, la valeur modale a été substituée à l'information manquante.

7.4.5 La période de pose

Pour un même matériau, il est possible de distinguer des plages temporelles au sein desquelles la technologie de fabrication ou d'installation des tuyaux peut être considérée comme homogène. L'exemple de la fonte grise discuté plus loin au chapitre 8 illustre bien l'impact de la période de pose sur la fiabilité des conduites, avec en particulier les taux de défaillance élevés observés pour les conduites posées durant le boom de la reconstruction après la seconde guerre mondiale.

Il est pertinent d'effectuer des regroupements d'années de pose et d'introduire leurs indicatrices dans le modèle. Il faut cependant veiller à ne considérer qu'un petit nombre de périodes larges afin de ne pas masquer l'effet de l'âge. La figure 7.3 porte pour l'ensemble des matériaux la distribution des années de pose groupées en quinquennats.

7.4.6 Le type d'encaissant

Nous pouvons distinguer, compte tenu des tailles d'échantillons disponibles les cinq catégories suivantes :

- sol compacté,
- sol naturel,
- appui ou remblai béton,
- tubage,
- hors sol.

La catégorie « sol compacté » est la plus fréquente quel que soit le matériau ; elle résulte d'une pose en tranchée ouverte, sans constitution d'un lit de pose particulier, suivie d'un remblayage avec compactage en couches successives du remblai. La catégorie « sol naturel » concerne principalement le cas (peu fréquent) des canalisations en PEHD et en béton âme-tôle posées avec une technique dite « sans tranchée » (micro-tunnelage par exemple). La catégorie « appui ou remblai béton » se rencontre avec tous les matériaux étudiés, et résulte d'un installation en tranchée ouverte avec constitution d'un lit de pose en béton, voire d'un remblayage complet avec du béton. La catégorie « tubage » est très minoritaire, et concerne des canalisations en fonte ductile et en PEHD posées à l'intérieur d'une canalisation préexistante (le plus souvent

Année	Longueur réseau (km)	Longueur mise hors service (km)	% de mise hors service	Nb de défaillances	Taux de défaillances
1995	7 871.7	20.3	0.258	732	0.093
1996	7 926.7	23.4	0.295	1 429	0.180
1997	7 981.0	31.0	0.388	1 384	0.173
1998	8 020.6	29.8	0.372	1 098	0.137
1999	8 056.4	28.5	0.354	1 053	0.131
2000	8 097.2	26.4	0.326	808	0.100
2001	8 146.0	28.1	0.345	1 153	0.142
2002	8 177.5	27.4	0.336	1 060	0.130
2003	8 218.4	29.3	0.356	1 454	0.177
2004	8 249.0	32.1	0.389	1 157	0.140
2005	8 279.4	34.1	0.412	1 391	0.168
2006	8 272.2	20.4	0.246	1 033	0.125

TAB. 7.1 – Linéaire (km) mis hors service et défaillances annuellement observés tous matériaux confondus

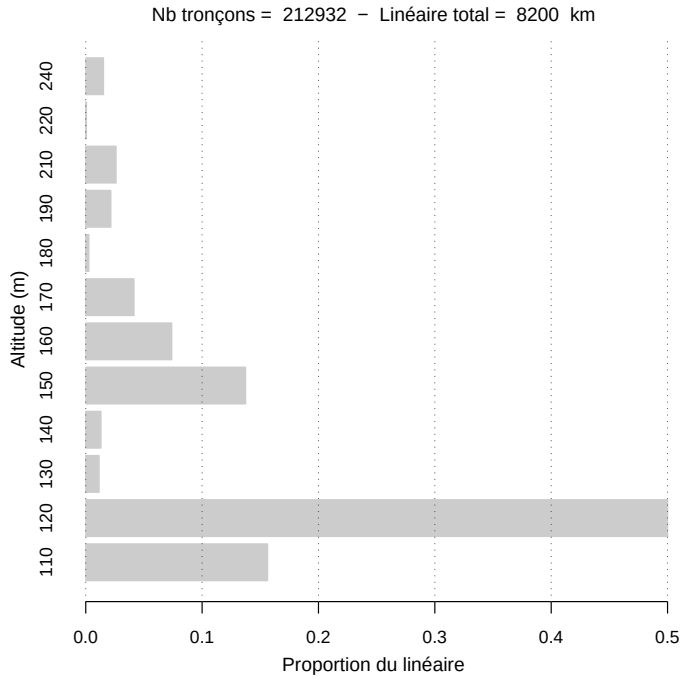


FIG. 7.2 – Distribution de l'altitude des canalisations (arrondie à la dizaine)

en fonte grise). Les canalisations posées hors sol sont tenues par des berceaux, des colliers, ou bien reposent sur dalle béton ou sur pieux. Les indicatrices des catégories les plus consistantes sont utilisées comme covariables.

7.4.7 Le type de joints

Les modalités de cette caractéristique dépendent fortement du matériau. Les distributions correspondantes sont données plus loin dans les sections relatives aux modèles par matériaux. Les fuites aux joints représentent une fraction non négligeable des défaillances, de 5 % pour la fonte grise à 65 % pour le béton âme tôle.

7.4.8 Le type d'occupation du sol

Le type d'occupation du sol caractérise l'environnement proche de la canalisation. Pour les canalisations enterrées, cela renseigne sur le risque d'exposition aux vibrations et contraintes de charges imposées par le trafic de véhicules au dessus du tronçon. Le tableau 7.2 porte les proportions du réseau concernées par ces emplacements. Dans la suite et pour des raisons de consistance, les types « parking ou esplanade », « voie ferrée », « transversal à une voie », « place » et « périphérie d'une place » sont regroupés avec « chaussée », les types « accotement » et « îlot central » avec « trottoir » ; les types restants (y compris le type inconnu, heureusement très peu représenté), censés être relatifs à une exposition très faible au trafic sont regroupés dans la catégorie « hors trafic ».

7.4.9 Les valeurs manquantes

Ce problème concerne essentiellement les types de joints et d'occupation du sol pour des canalisations remplacées au cours de la fenêtre d'observation. Selon le matériau, les valeurs manquantes ont été remplacées par le type dominant d'occupation du sol, et par le type dominant de joints compte tenu de la période de pose.

7.5 La sélection des covariables

La méthode retenue pour sélectionner le jeu de covariables pertinentes consiste à caler un modèle avec l'ensemble des covariables disponibles, puis à caler un nouveau modèle après élimination de la moins significative parmi les covariables non significatives, pour un seuil d'erreur de première espèce que nous avons fixé à 0.2 (*i.e.* 20 % de risque de se tromper en rejetant à tort l'hypothèse nulle « $\beta_j = 0$ »), et à itérer ce processus d'élimination descendante jusqu'à obtention d'un jeu de covariables toutes significatives.

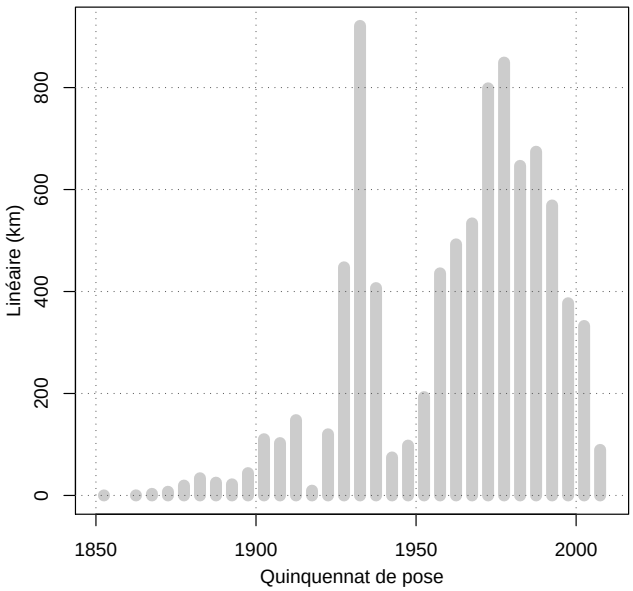


FIG. 7.3 – Distribution des quinquennats de pose

Occupation du sol	Effectif	% en effectif	Linéaire (km)	% en linéaire
Chaussée	122 339	57.38	4 407.3	53.67
Trottoir	70 683	33.15	3 245	39.52
Propriété privée	2 475	1.16	119.7	1.46
Parking ou esplanade	2 760	1.29	119.2	1.45
Transversal à une voie	7 074	3.32	73.3	0.89
Chemin	768	0.36	62.4	0.76
Voie piétonne	768	0.36	39.4	0.48
Espace vert	784	0.37	33.6	0.41
Ilot central	1 861	0.87	29.6	0.36
Accotement	353	0.17	15.8	0.19
Périphérie d'une place	871	0.41	15.7	0.19
Place	780	0.37	14.1	0.17
En ouvrage d'art	412	0.19	12.7	0.15
Autre	717	0.34	12.7	0.15
Inconnu	151	0.07	6.4	0.08
Voie ferrée	218	0.1	4.1	0.05
Position verticale	201	0.09	1	0.01
Total	213 215	100.00	8 212.2	100.00

TAB. 7.2 – Distribution des types d'occupation du sol

Chapitre 8

Le modèle fonte grise

La fonte grise est le matériau le plus anciennement posé en réseaux de distribution d'eau potable. Elle a été fabriquée jusqu'en 1965, date à laquelle la fonte ductile lui a été substituée. La pose de fonte grise a cependant pu perdurer jusqu'au début des années 70, en raison des stocks constitués par les services des eaux. Le tableau 8.1 porte la distribution des périodes de pose ; Les deux grosses vagues d'équipement des années 30 et 60 sont bien marquées.

Année de pose	Effectif	% en effectif	Linéaire (km)	% en linéaire
1850-1890	1 268	1.63	70.6	2.00
1891-1905	3 045	3.92	152.5	4.33
1906-1915	3 864	4.98	193.8	5.50
1916-1925	3 338	4.30	156.1	4.43
1926-1935	26 088	33.61	1269.0	35.99
1936-1945	6 990	9.01	331.7	9.41
1946-1955	6 929	8.93	313.5	8.89
1956-1965	18 458	23.78	739.6	20.98
1966-1971	7 634	9.84	299.2	8.49
Total	77 614	100.00	3 526.1	100.00

TAB. 8.1 – Distribution des classes d'années de pose des canalisations en fonte grise

8.1 Les défaillances

La répartition des défaillances par type est portée au tableau 8.2. En considération de la remarque 7.1, il serait hasardeux de rechercher sur ces données un lien entre le type de défaillance et la probabilité de mise hors service.

La figure 8.1 porte, en trait grisé continu, le graphe du taux de défaillances empirique (estimateur de Nelson-Aalen lissé par un noyau d'Epanechnikov présenté en sous-section 7.2.1) en fonction de l'âge des canalisations en fonte grise. La relation n'est pas monotone, et suggère de distinguer les tranches d'âges :

Type de défaillance	% en effectif	% en linéaire
Rupture circulaire	66.72	66.51
Rupture longitudinale	11.61	11.11
Eclat	9.75	9.88
Piqûre	6.12	6.38
Joint	5.22	5.36
Autre	0.57	0.76
10 308 défaillances typées sur 11 300 observées		

TAB. 8.2 – Distribution des types de défaillances des canalisations en fonte grise

– 25-60 ans,
– 60-100 ans,
– 100-120 ans,
– 120-140 ans,

qui correspondent aux quatre « générations » définies par les périodes de pose :

– 1850-1889,
– 1890-1930,
– 1931-1945,
– 1946-1970.

La relation semble conforme au modèle puissance du temps pour la seconde génération, à l'exception des cohortes extrêmes, représentées par des linéaires très faibles et qui exhibent un comportement plus chaotique. Il convient de noter le mauvais comportement des fontes posées après la seconde guerre mondiale, bien que caractérisées par une épaisseur régulière due à la technique de moulage par centrifugation ; les deux générations les plus anciennes (1850-1889 et 1890-1930) étaient moulées par simple coulage, et donc plus irrégulières en épaisseur ; la génération 1931-1945 marque le début de la technique de moulage par centrifugation. La piètre qualité des fontes grises fabriquées pendant les années de forte activité de (re-)construction d'après 1945, tient sans doute pour une part à une moindre qualité des minerais, mais aussi à la tendance à produire des épaisseurs plus fines que lorsqu'on entendait compenser l'irrégularité d'épaisseur obtenue par moulage par une paroi globalement plus épaisse. La génération de fonte grise la plus ancienne semble présenter un taux de défaillances croissant avec l'âge jusqu'à 110 ans, puis au contraire décroissant nettement avec l'âge, comportement vraisemblablement explicable par le phénomène de la survie sélective. La remontée apparente du taux de défaillance empirique au delà de 135 ans est sans doute un artefact dû à la forte raréfaction du linéaire observé à cet âge.

Les tableaux 8.3 et 8.4 portent respectivement pour les fontes grises produites avant et après 1945 les distributions des nombres de défaillances observées par tronçon entre 1995 et 2006, et ce pour les tronçons encore en service fin 2006, les tronçons mis hors service et l'ensemble. Les défaillances concernent globalement moins de 14 % des tronçons. 5.46 % des tronçons sont mis hors service entre 1995 et 2006, ce taux étant beaucoup plus élevé, 18.22 %, pour les fontes grises les plus anciennes. Il apparaît que les mises hors service concernent statistiquement plus

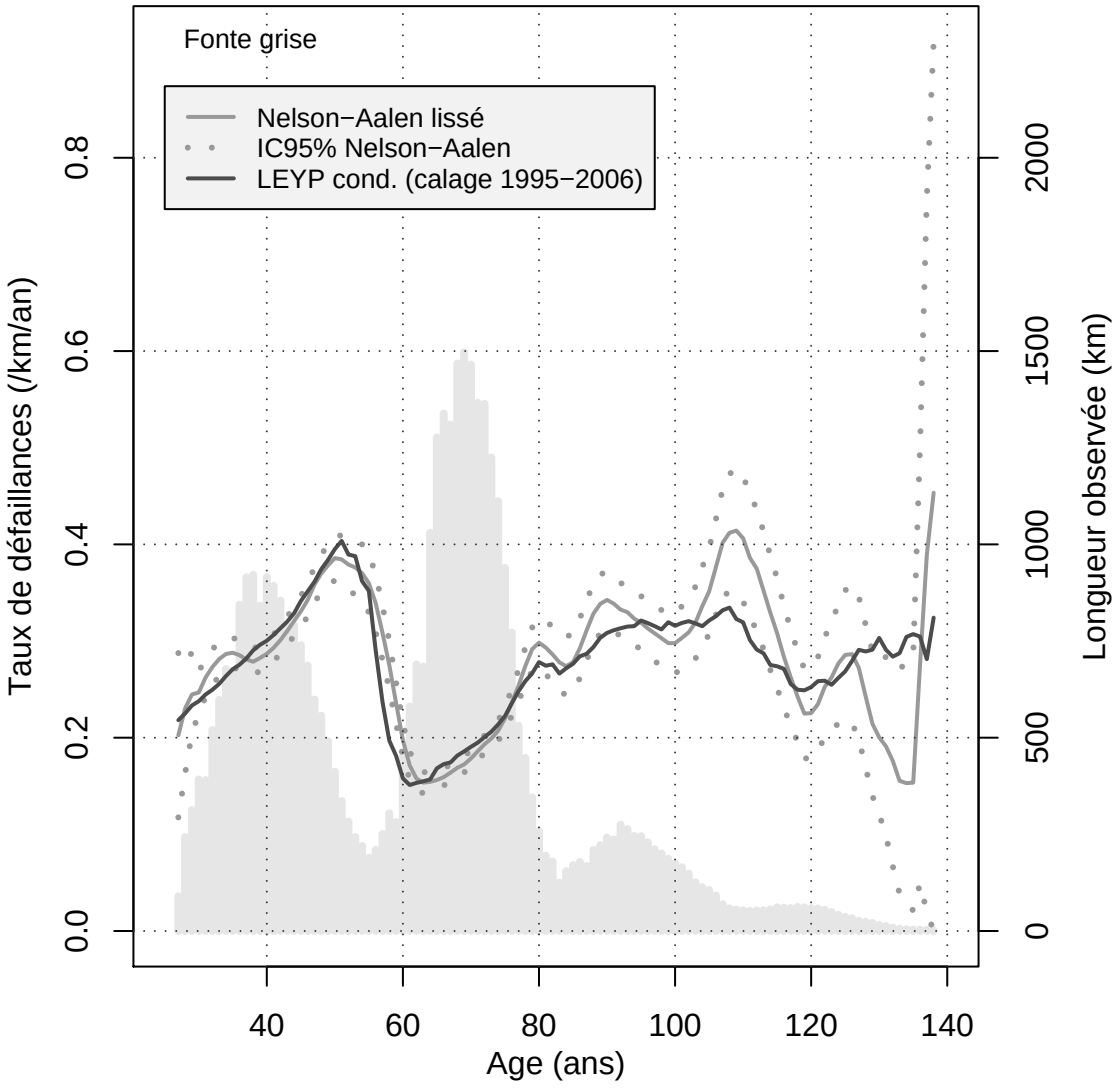


FIG. 8.1 – Taux de défaillance des tronçons en fonte grise selon leur âge

les tronçons ayant subi au moins une défaillances depuis 1995. Nous voyons ici à l'oeuvre le mécanisme de génération du biais de survie sélective. Ces données justifient le choix fait pour la forme analytique de $\varsigma(t)$ exposée en section 6.4.

Nb défaillances par tronçon	Tronçons non mis hors service		Tronçons mis hors service		Tous tronçons	
0	40 780	91.45	2 120	78.03	42 900	90.68
1	3 074	6.89	297	10.93	3 371	7.13
2	543	1.22	163	6.00	706	1.49
3	136	0.30	77	2.83	213	0.45
4	40	0.09	35	1.29	75	0.16
5	12	0.03	15	0.55	27	0.06
6	4	0.01	6	0.22	10	0.02
7	3	0.01	1	0.04	4	0.01
8	0	0.00	2	0.07	2	0.00
9	0	0.00	1	0.04	1	0.00
10	1	0.00	0	0.00	1	0.00
Total tronçons	44 593	100.00	2 717	100.00	47 310	100.00
Total défaillances	4 843		1 137		5 980	

TAB. 8.3 – Distribution des nombres de défaillances par tronçon en fonte grise posée avant 1945

Nb défaillances par tronçon	Tronçons non mis hors service		Tronçons mis hors service		Tous tronçons	
	Eff.	%	Eff.	%	Eff.	%
0	29 456	89.30	1136	74.59	30 592	88.65
1	2 734	8.29	214	14.05	2 948	8.54
2	580	1.76	97	6.37	677	1.96
3	152	0.46	44	2.89	196	0.57
4	39	0.12	17	1.12	56	0.16
5	15	0.05	9	0.59	24	0.07
6	6	0.02	6	0.39	12	0.03
7	2	0.01	0	0.00	2	0.01
Total tronçons	32 984	100.00	1 523	100.00	34 507	100.00
Total défaillances	4 631		689		5 320	

TAB. 8.4 – Distribution des nombres de défaillances par tronçon en fonte grise posée après 1945

Le tableau 8.5 porte les mises hors service et défaillances annuelles. Matériau majoritaire en linéaire sur le réseau, la fonte grise détient aussi le « record » des mises hors service (avec le

plus souvent remplacement par de la fonte ductile, et dans une moindre mesure par du PEHD) et du taux de défaillances. Les mises hors service sont en outre très ciblées sur les tronçons subissant des défaillances répétées. Il est de plus vraisemblable que la pression de mise hors service sur les plus défaillants se soit accrue au cours de la fenêtre d'observation, puisque se trouvant facilitée par la mise en place de l'archivage électronique des données de maintenance et leur couplage avec le SIG.

Année	Longueur réseau (km)	Longueur mise hors service (km)	% de mise hors service	Nb de défaillances	Taux de défaillance
1995	3 853.8	16.2	0.421	584	0.152
1996	3 837.6	21.9	0.571	1 163	0.303
1997	3 815.7	26.5	0.694	1 185	0.311
1998	3 789.2	27.8	0.733	935	0.247
1999	3 761.4	25.6	0.681	888	0.236
2000	3 735.8	23.6	0.631	671	0.180
2001	3 712.2	24.0	0.648	972	0.262
2002	3 688.2	25.5	0.693	879	0.238
2003	3 662.6	26.2	0.716	1 164	0.318
2004	3 636.4	29.7	0.817	948	0.261
2005	3 606.7	27.9	0.774	1 086	0.301
2006	3 578.8	16.6	0.464	825	0.231

TAB. 8.5 – Linéaire (km) mis hors service et défaillances annuellement observés pour les tronçons en fonte grise

8.2 Les covariables

La figure 8.2 porte la fonction de répartition de la longueur des tronçons en fonte grise ; la queue de distribution (derniers 10 %) est omise, car la distribution est nettement en « i » : 40 % des tronçons mesurent moins de 10 m , et 50 % moins de 20 m.

Les valeurs de diamètre sont regroupées en sept classes, dénommées d'après le diamètre majoritaire, et sa distribution est portée au tableau 8.6. Les diamètres compris entre 80 et 150 mm représentent 83 % du linéaire. Les diamètres 300 mm et plus ne concernent pas des canalisations de distribution, mais de transport de l'eau entre un point de production et un point de stockage ; ces *feeders* ne sont pas fragilisés par la présence de piquages de branchements particuliers, et ont une paroi d'une épaisseur qui leur permet de supporter des pressions élevées ; leur taux de défaillance est en conséquence toujours très bas.

La distribution des types de joints est portée au tableau 8.7. La population se partage grossièrement en deux tiers de joints coulés (technique ancienne du matage au plomb) et un tiers de joints express non vissés (technique prévalant après la seconde guerre mondiale).

Comme en témoigne le tableau 8.8, la profondeur de pose est peu variable, essentiellement située entre 1.10 et 1.20 m.

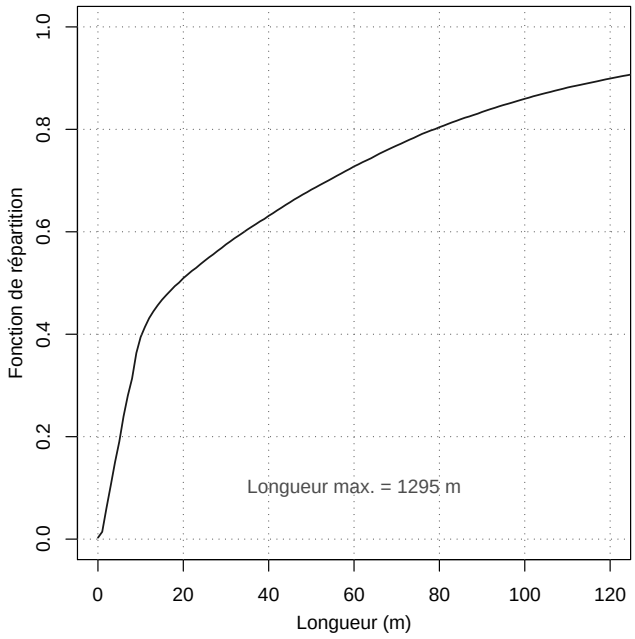


FIG. 8.2 – Fonction de répartition de la longueur des tronçons en fonte grise

Diamètre (mm)	Effectif	% en effectif	Linéaire (km)	% en linéaire
60	4 685	6.04	236.5	6.71
80	11 116	14.32	561.2	15.92
100	43 657	56.25	1 933.1	54.82
150	10 474	13.49	433.1	12.28
200-250	5 840	7.52	262.5	7.45
300-800	1 842	2.37	99.6	2.83
Total	77614	100.00	3526.1	100.00

TAB. 8.6 – Distribution des classes de diamètres des canalisations en fonte grise

Joints	Effectif	% en effectif	Linéaire (km)	% en linéaire
Coulé	46 266	59.61	2 250.6	63.83
Express non vissé	30 645	39.48	1 242.6	35.24
Gibault	689	0.89	32.4	0.92
Autre	14	0.02	0.6	0.02
Total	77 614	100.00	3 526.1	100.00

TAB. 8.7 – Distribution des types de joints sur canalisations en fonte grise

Profondeur (m)	Effectif	% en effectif	Linéaire (km)	% en linéaire
0.0-0.8	135	0.17	5.4	0.15
0.9-1.0	3 621	4.67	235.3	6.67
1.1-1.2	73 659	94.9	3 273.4	92.83
1.3-2.0	65	0.08	3	0.09
2.1-4.0	52	0.07	2.8	0.08
4.1-50.8	48	0.06	5.1	0.14
Hors Sol	34	0.04	1	0.03
Total	77 614	100.00	3 526.1	100.00

TAB. 8.8 – Distribution des profondeurs de pose des canalisations en fonte grise

Le tableau 8.9 montre que l'occupation du sol au dessus des canalisations se partage essentiellement entre la chaussée et le trottoir.

Occupation du sol	Effectif	% en effectif	Linéaire (km)	% en linéaire
Trottoir	31 786	40.95	1 892.1	53.66
Chaussée	44 945	57.91	1 580.6	44.83
Autre	883	1.14	53.4	1.51
Total	77 614	100.00	3 526.1	100.00

TAB. 8.9 – Distribution des types d'occupation du sol au dessus des canalisations en fonte grise

La technique de pose est quasi-exclusivement (99.91 % du linéaire) de type traditionnel. Le remblai en sol compacté représente 99.83 % du linéaire. Le tableau 8.10 porte la distribution de l'altitude des conduites.

Altitude	Effectif	% en effectif	Linéaire (km)	% en linéaire
105-114	11 169	14.39	589.2	16.71
115-124	38 785	49.97	1 683.3	47.74
125-154	15 500	19.97	693.6	19.67
155-174	8 091	10.42	364.2	10.33
175-244	4 069	5.24	195.9	5.55
Total	77 614	100.00	3 526.1	100.00

TAB. 8.10 – Distribution de l'altitude des canalisations en fonte grise

8.3 Diagnostic de la qualité d'ajustement du modèle

Les valeurs des paramètres du modèle ζ-LEYP calées sur les défaillances observées du 01/01/1995 au 31/12/2006 sont portées au tableau 8.11, assorties de leur écart type d'estimation, de la valeur prise en référence pour le test d'hypothèse nulle, de la statistique de chi carré et de sa probabilité de dépassement (*cf.* 5.2.2).

La qualité d'ajustement du modèle aux données peut être diagnostiquée graphiquement en examinant la figure 8.1 qui superpose les taux de défaillance théorique et empirique selon l'âge. La qualité d'ajustement est globalement satisfaisante, mais tend à se dégrader au delà de 100 ans, alors que les linéaires observés s'amenuisent fortement ; la courbe du taux de défaillance théorique sort cependant peu de l'intervalle de confiance du taux empirique lissé.

Libellé	Valeur estimée	Ecart type	Référence	χ^2	Pr > χ^2
Alpha	+1.9080e+00	6.4032e-02	0.0	2.6595e+03	0.000000
Delta	+1.1838e+00	6.2498e-02	1.0	1.0932e+04	0.000000
Zeta0	-3.0855e+00	9.7316e-02	$-\infty$	5.1230e+04	0.000000
Zeta1	+1.0722e-02	1.3890e-03	0.0	4.4801e+02	0.000000
Intercept	-7.5783e+00	2.6745e-01	0.0	8.0288e+02	0.000000
ln(Longueur)	+5.5189e-01	9.2245e-03	0.0	3.5795e+03	0.000000
Diamètre	-1.7020e-03	1.6044e-04	0.0	1.1253e+02	0.000000
Pose 1850-1889	-7.5021e-01	7.3112e-02	0.0	1.0529e+02	0.000000
Pose 1890-1930	-5.1766e-01	3.9072e-02	0.0	1.7554e+02	0.000000
Pose 1931-1945	-6.7031e-01	3.3792e-02	0.0	3.9347e+02	0.000000
Sous chaussée	+8.1884e-02	1.3764e-02	0.0	3.5393e+01	0.000000

TAB. 8.11 – Fonte grise - Paramètres du ζ-LEYP calés sur les observations du 01/01/1995 au 31/12/2006

La comparaison portée au tableau 8.12 des nombres totaux de défaillances, prédit et observé, pour la période de calage 1995 à 2006, montre une tendance significative à la surestimation de 4.7 %.

Période	Nb tronçons	Tot. obs.	Tot. préd.	IC95
[01/01/1995, 31/12/2006]	81 824	11 274	11 809.229	[11 521.871, 12 096.587]

TAB. 8.12 – Fonte grise - Comparaison des totaux observés et prédits

8.4 Probabilité de maintien en service après défaillance

La mise en oeuvre de la correction du biais de survie sélective telle qu'exposée au chapitre 6, montre une baisse très hautement significative avec l'âge de la probabilité de maintien en service après défaillance ; l'effet des paramètres ζ_0 et ζ_1 est visualisé à la figure 8.3.

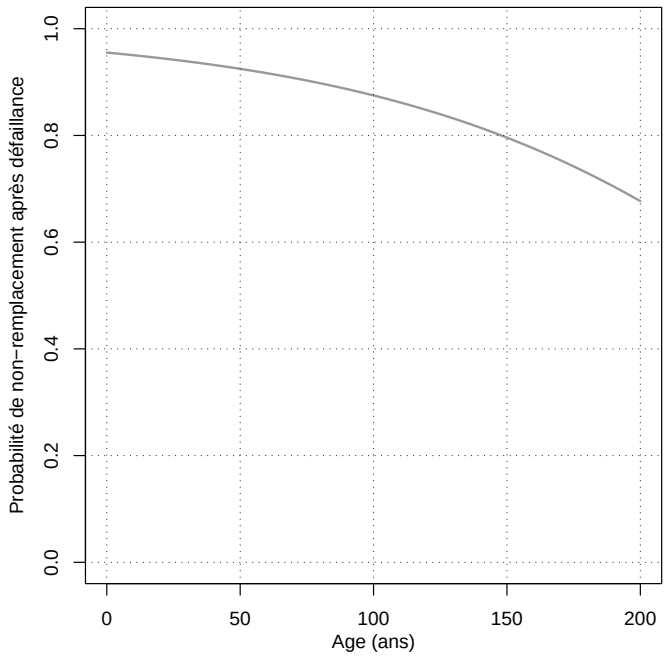


FIG. 8.3 – Probabilité de maintien en service après défaillance des canalisations en fonte grise

8.5 Effet des covariables

Avec un paramètre α proche de 2, la fonte grise manifeste une nette tendance des défaillances à s'accumuler sur les mêmes tronçons, comportement donc très éloigné du NHPP. L'effet de l'âge est plus discret, avec un δ un peu inférieur à 1.2 ; ce résultat est cependant compatible avec le fait que des fontes grises installées en environnement peu stressant puissent sembler quasiment « immortelles », à l'image de celles qui alimentent les grandes eaux du Château de Versailles depuis 1680.

Le tableau 8.13 porte les risques relatifs déduits des paramètres du modèle, portés au tableau 8.11, assortis de leur intervalle de confiance à 95 %. L'effet de la longueur du tronçon apparaît très proche de celui observé habituellement dans ce type d'étude, à savoir un accroissement du risque proportionnel à la racine carrée de la longueur du tronçon (*cf.* [Le Gat et Eisenbeis, 2000]) ; ainsi, un accroissement de la longueur de la conduite de 50 % se traduit théoriquement par un accroissement du risque instantané de défaillance de 25 %. L'effet proportionnel du diamètre, pris ici comme un facteur de risque continu, est très hautement significatif ; un accroissement de 100 mm du diamètre réduit le risque de 16 %. Les générations technologiques de fonte grise sont prises en compte par les variables indicatrices des périodes de pose 1850-1889, 1890-1930 et 1931-1945 (la période 1946-1970 est prise en référence) ; la troisième est caractérisée par l'innovation du moulage par centrifugation au début des années 30, celle de référence par les fabrications d'après-guerre, massives et à coût très contraint. Par rapport à la période de référence, les périodes 1850-1889, 1890-1930 et 1931-1945 manifestent des gains de fiabilité, avec des réductions du risque de 53 %, 40 % et 49 %, qui expliquent bien la première « bosse » observée autour d'un âge de 50 ans (référence) à la figure 8.1. L'installation sous chaussée semble accroître de façon très hautement significative le risque de 9 %, par rapport à

une situation de référence très majoritairement sous trottoir.

Covariable	Risque relatif	IC95 %
Longueur (×1.5)	1.251	[1.242,1.260]
Diamètre (+ 100 mm)	0.843	[0.817,0.870]
Pose 1850-1889	0.472	[0.409,0.545]
Pose 1890-1930	0.596	[0.552,0.643]
Pose 1931-1945	0.512	[0.479,0.547]
Sous chaussée	1.085	[1.056,1.115]

TAB. 8.13 – Risques relatifs associés aux covariables du modèle fonte grise

8.6 Performance prédictive des modèles ζ-LEYP et NHPP

Comme expliqué à la section 5.5, l'étude de la performance prédictive du modèle peut s'effectuer par validation, en calant les paramètres sur les neuf premières années de la chronique des défaillances, puis en comparant les prédictions du modèle aux défaillances réellement observées sur les trois dernières années de la chronique. Nous avons de plus choisi de faire cet exercice en parallèle pour le modèle ζ-LEYP et pour le modèle NHPP, auquel sa très large utilisation dans l'étude pratique des événements répétés confère le statut de « modèle de référence ».

8.6.1 Calage des modèles ζ-LEYP et NHPP

Le tableau 8.14 porte les valeurs des paramètres du modèle ζ-LEYP calées sur les défaillances observées du 01/01/1995 au 31/12/2003, assorties de leur écart type d'estimation, de la valeur prise en référence pour le test d'hypothèse nulle, de la statistique de chi carré et de sa probabilité de dépassement (*cf.* 5.2.2).

De même, le tableau 8.15 porte les valeurs des paramètres du modèle NHPP. La comparaison avec les paramètres du ζ-LEYP montre que le δ du NHPP est bien plus élevé, ce qui est logique puisque toute la dynamique des défaillances repose sur lui, alors qu'elle est partagée avec α dans le cas du ζ-LEYP. Les signes des β sont les mêmes entre les deux modèles, alors que les $|\beta|$ du NHPP sont tous sensiblement plus élevés ; le rapport entre $|\beta_j|$, $j \geqslant 1$ et $|\beta_0|$ reste cependant à peu près le même pour un j donné entre les deux modèles.

8.6.2 Diagnostic de la qualité d'ajustement des modèles ζ-LEYP et NHPP

Le tableau 8.16 porte, pour les modèles ζ-LEYP et NHPP, la comparaison des nombres de défaillances observés et prédits sur les fenêtres de calage (du 01/01/1995 au 31/12/2003) et de validation (du 01/01/2004 au 31/12/2006).

Le modèle ζ-LEYP manifeste une tendance légère mais significative à surestimer de 3.3 % les défaillances dans la fenêtre de calage. Cela peut être dû à l'incapacité du modèle dans sa

Libellé	Valeur estimée	Ecart type	Référence	χ^2	Pr > χ^2
Alpha	+2.1266e+00	8.3289e-02	0.0	2.0363e+03	0.000000
Delta	+1.1770e+00	6.7199e-02	1.0	7.9887e+03	0.000000
Zeta0	-3.1166e+00	1.0684e-01	$-\infty$	4.1046e+04	0.000000
Zeta1	+1.2425e-02	1.5207e-03	0.0	4.7203e+02	0.000000
Intercept	-7.5347e+00	2.8082e-01	0.0	7.1989e+02	0.000000
ln(Longueur)	+5.3107e-01	1.0107e-02	0.0	2.7607e+03	0.000000
Diamètre	-1.4347e-03	1.7330e-04	0.0	6.8538e+01	0.000000
Pose 1850-1889	-6.9679e-01	8.0200e-02	0.0	7.5485e+01	0.000000
Pose 1890-1930	-5.1185e-01	4.2630e-02	0.0	1.4417e+02	0.000000
Pose 1931-1945	-6.8133e-01	3.7270e-02	0.0	3.3419e+02	0.000000
Sous chaussée	+9.8543e-02	1.5397e-02	0.0	4.0961e+01	0.000000

TAB. 8.14 – Fonte grise - Paramètres du ζ-LEYP calés sur les observations du 01/01/1995 au 31/12/2003

Libellé	Valeur estimée	Ecart type	Référence	χ^2	Pr > χ^2
Delta	+1.9812e+00	7.6512e-02	1.0	6.1525e+04	0.000000
Intercept	-1.1502e+01	3.2346e-01	0.0	1.2645e+03	0.000000
ln(Longueur)	+8.4488e-01	1.0731e-02	0.0	6.1988e+03	0.000000
Diamètre	-2.4152e-03	2.4142e-04	0.0	1.0009e+02	0.000000
Pose 1850-1889	-1.2527e+00	1.0874e-01	0.0	1.3271e+02	0.000000
Pose 1890-1930	-8.8300e-01	5.7957e-02	0.0	2.3212e+02	0.000000
Pose 1931-1945	-1.1712e+00	4.8069e-02	0.0	5.9360e+02	0.000000
Sous chaussée	+1.7377e-01	2.2462e-02	0.0	5.9849e+01	0.000000

TAB. 8.15 – Fonte grise - Paramètres du NHPP calés sur les observations du 01/01/1995 au 31/12/2003

forme actuelle à prendre en compte l'effet des campagnes de recherche active de fuite, qui génèrent des réparations de défaillances avant que ces dernières ne se manifestent d'elles même ; cet effet d'anticipation des défaillances gêne le calage du modèle car l'effort de recherche de fuites n'est pas uniforme dans le temps et, en particulier, peut voir son intensité varier notablement d'une année sur l'autre. Sur la fenêtre de validation le biais de prédiction est un peu inférieur à 2 % et non significatif.

Le tableau 8.16 traduit la capacité du modèle NHPP à prédire sans biais le nombre global de défaillances dans la fenêtre de calage, mais à surestimer légèrement, mais significativement, les prédictions dans la fenêtre de validation.

Nous pouvons donc considérer que les deux modèles ζ-LEYP et NHPP ont tous deux une qualité d'ajustement satisfaisante sur les périodes de calage et de validation choisies.

Période	Nb tronçons	Nb défaillances		IC95
1995 – 2003	81 824	ζ-LEYP	8 703.4	[8 469.5, 8 937.3]
		NHPP	8 420.5	[8 240.6, 8 600.3]
		Observation	8 421	
2004 – 2006	78 618	ζ-LEYP	2 910.7	[2 798.7, 3 022.7]
		NHPP	2 984.9	[2 877.8, 3 092.0]
		Observation	2 854	

Tab. 8.16 – Comparaison des nombres de défaillances observés et prédits pour la fonte grise avec les modèle ζ-LEYP et NHPP

8.6.3 Validation

Les prédictions de défaillances des canalisations en fonte grise entre le 01/01/2004 et le 31/12/2006, calculées grâce au modèle calé sur les défaillances observées du 01/01/1995 au 31/12/2003, sont comparées aux défaillances réellement observées sur la même période aux fins de validation du pouvoir prédictif des modèles ζ-LEYP et NHPP. La courbe de performance prédictive obtenue conformément à la méthode présentée en sous-section 5.5.1 est portée en figure 8.4. Les résultats un peu décevants (aire sous la courbe de 0.642) en détection des tronçons les plus à risque, mais cependant supérieurs au NHPP (aire sous la courbe de 0.591), tiennent sans doute à la focalisation des efforts de renouvellement de canalisations consentis lors de la décennie écoulée sur ce matériau par le gestionnaire du réseau ; il est vraisemblable que la poursuite de cette stratégie aboutisse en quelques années à l'élimination quasi-totale des canalisations susceptibles de manifester des défaillances répétées, rendant ainsi inutile l'utilisation de ce type de modèle pour prioriser les renouvellements annuels sur ce matériau ; le modèle gardera cependant tout son intérêt pour simuler l'effet à long terme de différents efforts de renouvellement sur la fiabilité globale d'une population de tronçons.

Aux conditions actuelles le modèle ζ-LEYP permet cependant encore d'orienter les renouvellements sur les tronçons les plus à risque, comme en témoigne le tableau 8.17 portant les proportions de défaillances qui auraient été évitées si la proportion correspondante du linéaire avait été renouvelée, en supposant le renouvellement prioritaire des tronçons que le modèle

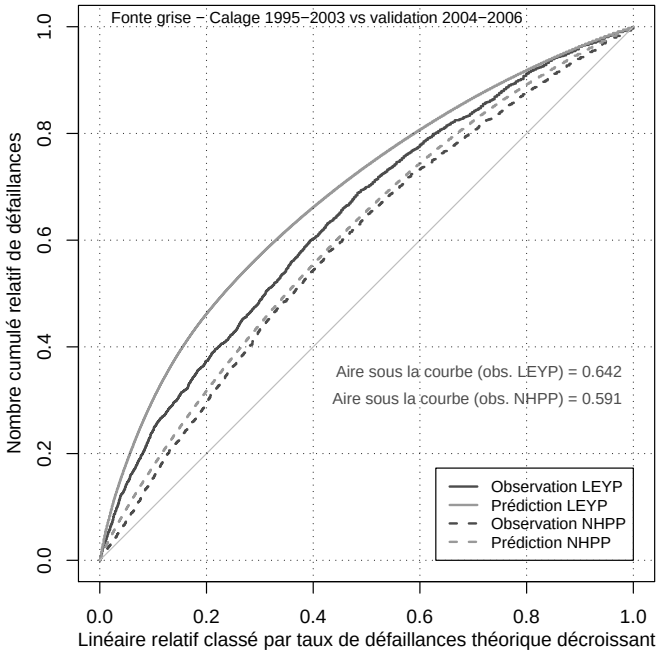


FIG. 8.4 – Performance prédictive du modèle fonte grise

classe comme étant les plus à risque. La performance du NHPP apparaît très sensiblement en retrait sur celle du ζ-LEYP.

% linéaire renouvelé		0.1	0.5	1.0	5.0
% défaillances évitables avec	ζ-LEYP	0.4	2.3	3.6	14.1
	NHPP	0.2	1.1	1.7	7.4

TAB. 8.17 – Proportions de défaillances évitables pour une proportion donnée de linéaire de fonte grise renouvelée selon les modèles ζ-LEYP et NHPP

La figure 8.5 permet de visualiser plus globalement, pour les 5 % du linéaire les plus à risque (respectivement au sens du ζ-LEYP et du NHPP), la supériorité de la performance prédictive du ζ-LEYP sur le modèle de référence.

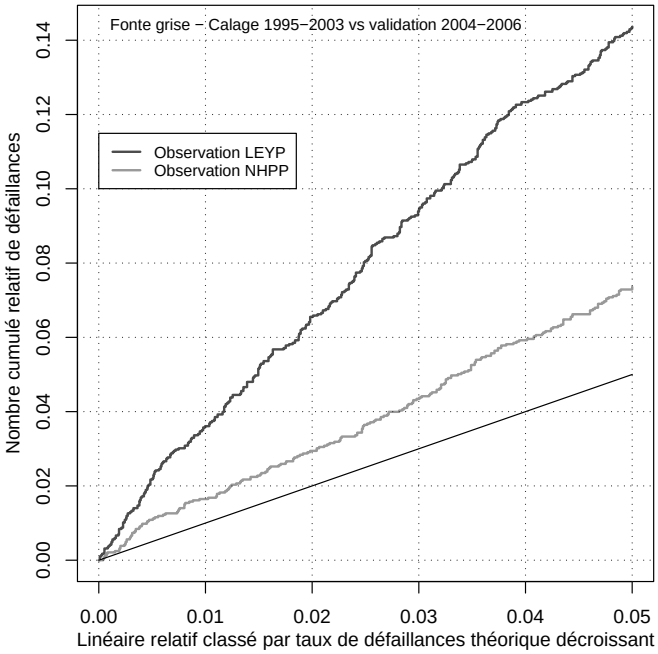


FIG. 8.5 – Performance prédictive du modèle fonte grise pour les 5 % du linéaire les plus à risque

Chapitre 9

Le modèle fonte ductile

Utilisée depuis le milieu des années 60, la fonte ductile se substitue progressivement à la fonte grise au gré des renouvellements des tronçons. Le tableau 9.1 porte la distribution des périodes (quinquennales) de pose.

Année de pose	Effectif	% en effectif	Linéaire (km)	% en linéaire
1965-1970	3 032	3.55	120.9	3.98
1971-1975	17 636	20.63	683.1	22.49
1976-1980	18 467	21.60	694.2	22.86
1981-1985	15 202	17.78	534.8	17.61
1986-1990	15 150	17.72	491.7	16.19
1991-1995	9 454	11.06	297.3	9.79
1996-2000	3 929	4.60	125.3	4.13
2001-2006	2 625	3.07	89.8	2.96
Total	85 495	100.00	3 037.1	100.00

TAB. 9.1 – Distribution des quinquennats de pose des canalisations en fonte ductile

9.1 Les défaillances

La répartition des défaillances par type est portée au tableau 9.2. Cette répartition apparaît assez proche de celle de la fonte grise (tableau 8.2), avec cependant environ 5 % de moins pour les ruptures circulaires (la fonte ductile a une meilleure résistance mécanique), au profit des piqûres de corrosion et des fuites aux joints. En considération de la remarque 7.1, il serait hasardeux de rechercher sur ces données un lien entre le type de défaillance et la probabilité de mise hors service.

La figure 9.1 porte les graphes des taux de défaillance empirique (Nelson-Aalen lissé) et théorique (ζ-LEYP) en fonction de l'âge. Ce matériau est d'utilisation somme toute récente ; un linéaire de 200 km a été posé depuis 1995, dont la fenêtre d'observation permet de suivre les

Type de défaillance	% en effectif	% en linéaire
Rupture circulaire	61.15	60.62
Rupture longitudinale	11.60	11.75
Eclat	10.38	10.71
Piqûre	8.76	8.91
Joint	7.54	7.36
Autre	0.57	0.66
1 233 défaillances typées sur 1 408 observées		

TAB. 9.2 – Distribution des types de défaillances des canalisations en fonte ductile

défaillances depuis la pose. Les défaillances « juvéniles » (*i.e.* survenant dans les mois suivant la pose) n'apparaissent cependant pas nettement sur la figure 9.1, à cause du lissage pratiqué sur le taux de défaillance empirique. Bien que restant toujours modeste, le taux de défaillance augmente nettement comme une puissance de l'âge. Il est difficile de départager les effets du biais de survie sélective de ceux du très faible linéaire observé dans l'explication de l'amorce de baisse du taux de défaillances empirique au delà de 34 ans.

Le tableau 9.3 portant la distribution des nombres de défaillances observées par tronçon entre 1995 et 2006, et ce pour les tronçons encore en service fin 2006, les tronçons mis hors service et l'ensemble, laisse apparaître que les défaillances concernent globalement moins de 2 % des tronçons ; cela illustre bien le gain plus que sensible de fiabilité par rapport à la fonte grise ; seulement 0.53 % des tronçons sont d'ailleurs mis hors service entre 1995 et 2006. Il apparaît cependant que les mises hors service concernent statistiquement plus les tronçons ayant subi au moins une défaillance depuis 1995. Les données de défaillances sur fonte ductile ne sont sans doute donc pas indemnes d'un certain biais de survie sélective.

Nb défaillances par tronçon	Tronçons non mis hors service		Tronçons mis hors service		Tous tronçons	
	Eff.	%	Eff.	%	Eff.	%
0	84 371	98.64	416	91.83	84 787	98.61
1	1 018	1.19	17	3.75	1 035	1.20
2	117	0.14	10	2.21	127	0.15
3	21	0.02	5	1.10	26	0.03
4	5	0.01	4	0.88	9	0.01
5	0	0.00	1	0.22	1	0.00
Total tronçons	85 532	100.00	453	100.00	85 985	100.00
Total défaillances	1 335		73		1 408	

TAB. 9.3 – Distribution des nombres de défaillances par tronçon en fonte ductile

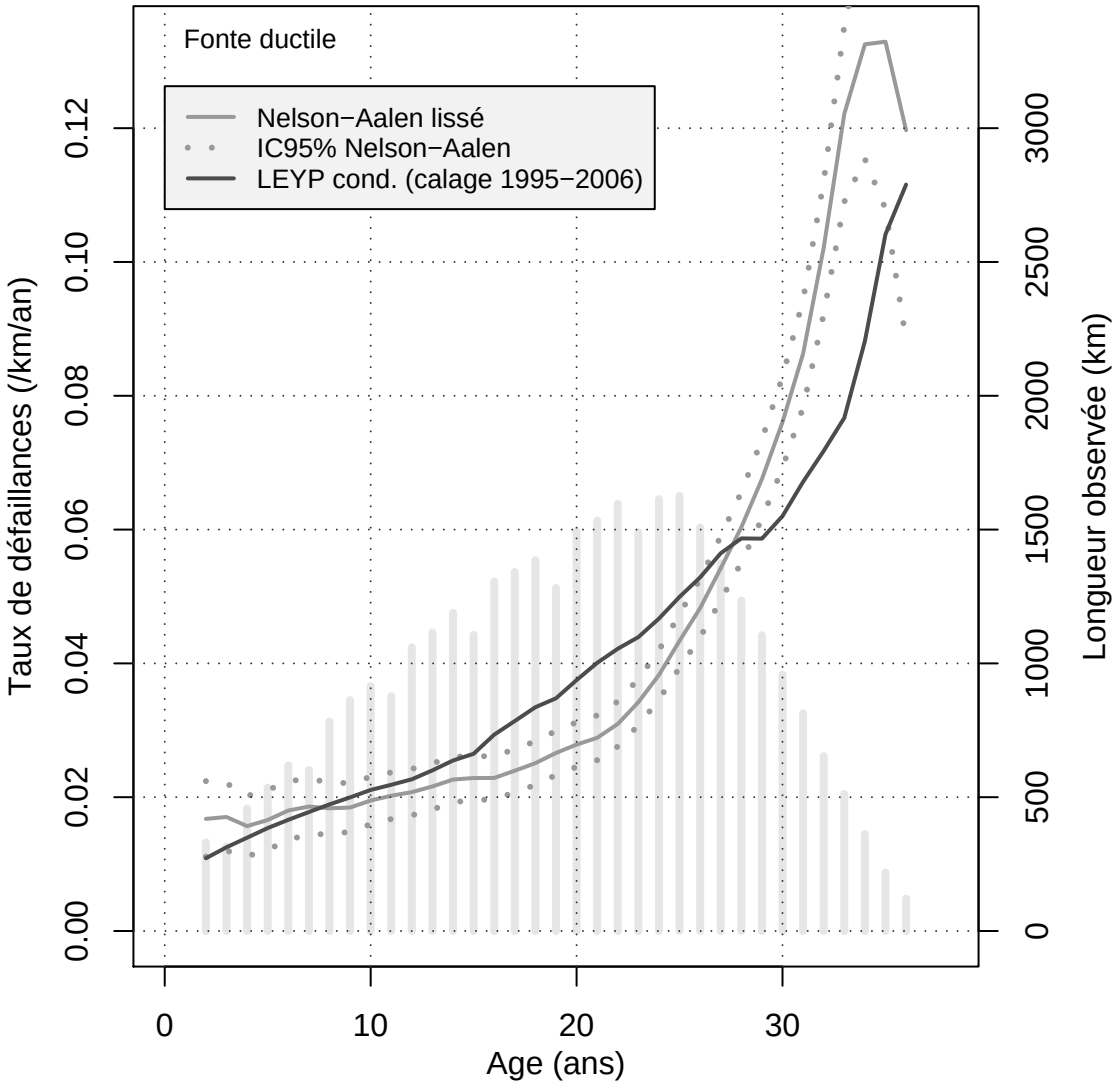


FIG. 9.1 – Taux de défaillance des tronçons en fonte ductile selon leur âge

Le tableau 9.4 porte les mises hors service et défaillances annuelles.

Année	Longueur réseau (km)	Longueur mise hors service (km)	% de mise hors service	Nb de défaillances	Taux de défaillances
1995	2 847.3	2.0	0.069	81	0.028
1996	2 882.8	1.1	0.039	157	0.054
1997	2 916.4	2.5	0.087	99	0.034
1998	2 936.1	1.7	0.059	102	0.035
1999	2 948.6	1.9	0.065	98	0.033
2000	2 963.9	1.4	0.047	80	0.027
2001	2 981.6	3.3	0.110	108	0.036
2002	2 993.0	0.9	0.030	107	0.036
2003	3 009.6	2.0	0.068	162	0.054
2004	3 023.0	1.3	0.044	114	0.038
2005	3 038.8	2.6	0.087	190	0.063
2006	3 042.1	2.9	0.097	110	0.036

TAB. 9.4 – Linéaire (km) mis hors service et défaillances annuellement observés pour les canalisations en fonte ductile

9.2 Les covariables

Les classes de diamètres, dont la distribution est portée au tableau 9.5, sont identiques à celles considérées pour la fonte grise, à l'exception de la quasi-absence du 60 mm, et de la faible présence du 80 mm (regroupés avec les 100 mm).

Diamètre (mm)	Effectif	% en effectif	Linéaire (km)	% en linéaire
080-100	44 580	52.14	1 469.0	48.37
125-150	16 219	18.97	594.5	19.57
200-250	15 756	18.43	589.7	19.42
300-800	8 940	10.46	384.0	12.64
Total	85 495	100.00	3 037.1	100.00

TAB. 9.5 – Distribution des classes de diamètres des canalisations en fonte ductile

La distribution des types de joints est portée au tableau 9.6. Le type automatique non vissé est très majoritaire.

La technique traditionnelle de pose concerne 99.2 % des tronçons et 99.5 % du linéaire. 98.6 % des tronçons, soit 99.3 % du linéaire, sont installés en remblai compacté. Le tableau 9.7

Joints	Effectif	% en effectif	Linéaire (km)	% en linéaire
Express non vissé	39 941	46.72	1 525.6	50.23
Automatique non vissé	44 277	51.79	1 471.5	48.45
Automatique vissé	645	0.75	24.6	0.81
Express vissé	285	0.33	12.7	0.42
Gibault	110	0.13	1.2	0.04
A bride	206	0.24	0.8	0.03
Autre	31	0.04	0.6	0.02
Total	85 495	100.00	3 037.1	100.00

TAB. 9.6 – Distribution des types de joints sur canalisations en fonte ductile

porte la distribution des altitudes des canalisations en fonte ductile ; la moitié se situe autour de 120 m.

Altitude (m)	Effectif	% en effectif	Linéaire (km)	% en linéaire
105-114	11 216	13.12	441.4	14.53
115-124	44 260	51.77	1 451.6	47.8
125-154	14 898	17.43	538.3	17.72
155-174	10 216	11.95	384.3	12.65
175-244	4 905	5.74	221.5	7.29
Total	85 495	100.00	3 037.1	100.00
Total	85 495	100.00	3 037.1	100.00

TAB. 9.7 – Distribution des altitudes des canalisations en fonte ductile

Comme en témoigne le tableau 9.8, la profondeur de pose est peu variable, très majoritairement entre 90 cm et 1 m.

9.3 Diagnostic de la qualité d'ajustement du modèle

Un premier modèle ζ-LEYP complet, (*i.e.* avec α, δ, ζ_0 et ζ_1 laissés libres) a été recherché en éliminant pas à pas de façon descendante les covariables non significatives au seuil de 20%. Il est cependant apparu que ζ_1 était peu significatif dans le modèle final, conséquence de l'emploi relativement récent de la fonte ductile, et donc du peu de recul quant à l'influence de l'âge sur la décision de renouvellement en cas de défaillance. Le modèle a donc été réajusté en fixant $\zeta_1 = 0$. Le tableau 9.9 porte les valeurs finalement estimées pour les paramètres, assorties de leur écart type d'estimation, de la valeur prise en référence pour le test d'hypothèse nulle, de la statistique de chi carré et de sa probabilité de dépassement (*cf.* 5.2.2). La probabilité de maintien

Profondeur	Effectif	% en effectif	Linéaire (km)	% en linéaire
0.0-0.8	283	0.33	11.3	0.37
0.9-1.0	81 141	94.91	2 874.0	94.63
1.1-1.2	3 032	3.55	115.5	3.80
1.3-2.0	209	0.24	10.4	0.34
2.1-4.0	205	0.24	9.8	0.32
4.1-28.1	180	0.21	11.5	0.38
Hors Sol	445	0.52	4.7	0.15
Total	85 495	100.00	3 037.1	100.00

TAB. 9.8 – Distribution des profondeurs de pose des canalisations en fonte ductile

en service suite à une défaillance est estimée à $\exp(-\exp(-3.6375)) = 0.9740$ avec un intervalle de confiance à 95% de $[0.9643, 0.9811]$.

Libellé	Valeur estimée	Ecart type	Référence	χ^2	Pr > χ^2
Alpha	+6.4851e+00	6.0863e-01	0.0	3.9609e+02	0.000000
Delta	+1.3527e+00	8.3217e-02	1.0	2.0363e+03	0.000000
Zeta0	-3.6375e+00	1.6426e-01	$-\infty$	1.8771e+03	0.000000
Zeta1	+0.0000e+00		0.0		
Intercept	-9.2417e+00	3.8606e-01	0.0	5.7306e+02	0.000000
ln(Longueur)	+6.0583e-01	2.2424e-02	0.0	7.2996e+02	0.000000
Diam. 100	+1.0979e+00	1.2638e-01	0.0	7.5470e+01	0.000000
Diam. 150	+1.0050e+00	1.3298e-01	0.0	5.7116e+01	0.000000
Diam. 200	+2.9392e-01	1.4893e-01	0.0	3.8950e+00	0.048429
Joint auto. non vissé	-3.7100e-01	7.0401e-02	0.0	2.7771e+01	0.000000
Technique tradition.	-1.1823e+00	2.6482e-01	0.0	1.9932e+01	0.000008
Sol compacté	-4.6956e-01	2.7883e-01	0.0	2.8359e+00	0.092180
Altitude 110	-2.0439e-01	7.5319e-02	0.0	7.3642e+00	0.006654
Prof. hors sol	-1.1287e+00	4.8786e-01	0.0	5.3531e+00	0.020686

TAB. 9.9 – Fonte ductile - Paramètres du ζ-LEYP calés sur les observations du 01/01/1995 au 31/12/2006

La valeur très basse de β_0 (-9.24), en comparaison de celle observée pour la fonte grise (-7.58), confirme l'opinion favorable qu'ont en général les fontainiers et les canalisateurs de la fiabilité de la fonte ductile. Cependant le facteur de vieillissement δ affiche dans notre étude une valeur élevée suggérant que ce matériau pourrait ne pas vieillir aussi bien que la fonte grise. La valeur très élevée du facteur de Yule α traduit la très forte agrégation des défaillances sur un faible % des tronçons, ce qui permet de tempérer ce pronostic un peu sombre ; le renouvellement rapide d'un petit nombre de tronçons très à risque devrait permettre d'assurer une bonne fiabilité pour la population ainsi sélectionnée.

La comparaison portée au tableau 9.10 des nombres totaux de défaillances, prédit et observé, pour la période de calage 1995 à 2006, montre une tendance peu marquée et non significative à la surestimation de 1.9 %.

Période	Nb tronçons	Tot. obs.	Tot. préd.	IC95
[01/01/1995, 31/12/2006]	85 947	1 404	1 430.469	[1 325.146, 1 535.793]

TAB. 9.10 – Fonte ductile - Comparaison des totaux observés et prédits

9.4 Effet des covariables

Le tableau 9.11 porte les risques relatifs afférents aux covariables conservées comme significatives. L'effet de la longueur du tronçon apparaît, comme pour la fonte grise, très proche de celui observé habituellement dans ce type d'étude ; ainsi, un accroissement de la longueur de la conduite de 50 % se traduit théoriquement par un accroissement du risque instantané de défaillance de 28 %. Les classes de diamètres ont un effet prépondérant, avec une tendance marquée défavorable aux diamètres faibles (100 et 150 mm). Le joint automatique non vissé paraît réduire le risque de défaillances d'environ 30% par rapport au joint express. La technique de pose traditionnelle semble réduire nettement (69 %) le risque par rapport à la pose en fourreau ou en tubage. L'encaissant en sol compacté semble réduire sensiblement le risque (37 %) par rapport à la présence d'un lit de pose en béton ; cette observation quelque peu contre intuitive s'explique sans doute par la propension du service gestionnaire à opter pour le lit de pose en béton dans les situations à hauts risques de stress mécanique pour les conduites. Les situations topographiques les plus basses (altitude autour de 110 m) semblent aussi réduire un peu le risque (18 %). Les canalisations situées hors sol présentent enfin un risque nettement moindre (68 %) relativement à celles qui sont enterrées.

Covariable	Risque relatif	IC95%
Longueur (×1.5)	1.278	[1.256,1.301]
Diamètre 100 mm	2.998	[2.340,3.841]
Diamètre 150 mm	2.732	[2.105,3.545]
Diamètre 200 mm	1.342	[1.002,1.796]
Joint Automatique Non Vissé	0.690	[0.601,0.792]
Pose traditionnelle	0.307	[0.182,0.515]
Sol compacté	0.625	[0.362,1.080]
Altitude 110 m	0.815	[0.703,0.945]
Hors sol	0.323	[0.124,0.842]

TAB. 9.11 – Risques relatifs associés aux covariables du modèle fonte ductile

9.5 Performance prédictive des modèles ζ-LEYP et NHPP

L'étude de la performance prédictive du modèle est effectuée par validation, en calant les paramètres sur les neuf premières années de la chronique des défaillances, puis en comparant les prédictions du modèle aux défaillances réellement observées sur les trois dernières années de la chronique. Nous avons mené cet exercice en parallèle pour le modèle ζ-LEYP et pour le modèle NHPP pris comme référence.

9.5.1 Calage des modèles ζ-LEYP et NHPP

Le tableau 9.12 porte les valeurs des paramètres du modèle ζ-LEYP calées sur les défaillances observées du 01/01/1995 au 31/12/2003, assorties de leur écart type d'estimation, de la valeur prise en référence pour le test d'hypothèse nulle, de la statistique de chi carré et de sa probabilité de dépassement (*cf.* 5.2.2).

Libellé	Valeur estimée	Ecart type	Référence	χ^2	Pr > χ^2
Alpha	+6.9295e+00	8.3431e-01	0.0	2.4324e+02	0.000000
Delta	+1.4233e+00	1.0060e-01	1.0	1.8632e+03	0.000000
Zeta0	-3.4958e+00	1.8245e-01	$-\infty$	1.5306e+03	0.000000
Zeta1	+0.0000e+00		0.0		
Intercept	-9.4798e+00	4.5462e-01	0.0	4.3481e+02	0.000000
ln(Longueur)	+6.1032e-01	2.6626e-02	0.0	5.2543e+02	0.000000
Diam. 100	+1.0567e+00	1.4640e-01	0.0	5.2100e+01	0.000000
Diam. 150	+9.3077e-01	1.5469e-01	0.0	3.6205e+01	0.000000
Diam. 200	+2.4505e-01	1.7387e-01	0.0	1.9863e+00	0.158725
Joint auto. non vissé	-3.5040e-01	8.6502e-02	0.0	1.6409e+01	0.000051
Technique tradition.	-1.3047e+00	2.7898e-01	0.0	2.1869e+01	0.000003
Sol compacté	-3.8326e-01	2.9670e-01	0.0	1.6686e+00	0.196451
Altitude 110	-1.2494e-01	8.5869e-02	0.0	2.1171e+00	0.145663
Prof. hors sol	-1.4591e+00	6.8302e-01	0.0	4.5638e+00	0.032654

Tab. 9.12 – Fonte ductile - Paramètres du ζ-LEYP calés sur les observations du 01/01/1995 au 31/12/2003

De même, le tableau 9.13 porte les valeurs des paramètres du modèle NHPP. La comparaison avec les paramètres du ζ-LEYP montre, à l'instar de la fonte grise, que le δ du NHPP est sensiblement plus élevé. Les signes des β sont les mêmes entre les deux modèles, alors que les $|\beta|$ du NHPP sont tous sensiblement plus élevés, de même que les rapports entre $|\beta_j|$, $j \geqslant 1$ et $|\beta_0|$. Il semble donc qu'ici le ζ-LEYP atténue sensiblement l'effet des covariables comparativement au NHPP.

9.5.2 Diagnostic de la qualité d'ajustement des modèles ζ-LEYP et NHPP

Le tableau 9.14 porte, pour le modèle ζ-LEYP, la comparaison des nombres de défaillances observés et prédits sur les fenêtres de calage et de validation. Le modèle manifeste une tendance

Libellé	Valeur estimée	Ecart type	Référence	χ^2	Pr > χ^2
Delta	+1.7643e+00	1.1041e-01	1.0	3.8277e+03	0.000000
Intercept	-1.0137e+01	5.0489e-01	0.0	4.0313e+02	0.000000
ln(Longueur)	+7.5338e-01	2.7996e-02	0.0	7.2413e+02	0.000000
Diam. 100	+1.3370e+00	1.5898e-01	0.0	7.0719e+01	0.000000
Diam. 150	+1.1876e+00	1.6932e-01	0.0	4.9189e+01	0.000000
Diam. 200	+3.4611e-01	1.8882e-01	0.0	3.3598e+00	0.066805
Joint auto. non vissé	-3.1743e-01	9.9726e-02	0.0	1.0132e+01	0.001457
Technique tradition.	-2.4823e+00	3.0095e-01	0.0	6.8033e+01	0.000000
Sol compacté	-3.5076e-01	3.2915e-01	0.0	1.1356e+00	0.286589
Altitude 110	-1.7711e-01	9.8452e-02	0.0	3.2362e+00	0.072027
Prof. hors sol	-2.4390e+00	7.2314e-01	0.0	1.1376e+01	0.000744

Tab. 9.13 – Fonte ductile - Paramètres du NHPP calés sur les observations du 01/01/1995 au 31/12/2003

légère non significative à surestimer de 2.5 % les défaillances dans la fenêtre de calage (du 01/01/1995 au 31/12/2003). Sur la fenêtre de validation (du 01/01/2004 au 31/12/2006) le biais de prédiction est de -3.7 % et non significatif.

Le tableau 9.14 traduit la capacité du modèle NHPP à prédire sans biais le nombre global de défaillances dans la fenêtre de calage, mais à sousestimer légèrement et non significativement de 5.4 % les prédictions dans la fenêtre de validation.

Nous pouvons donc considérer que les deux modèles ζ-LEYP et NHPP ont tous deux une qualité d'ajustement satisfaisante sur les périodes de calage et de validation choisies.

Période	Nb tronçons	Nb défaillances		IC95
1995 – 2003	84 731	ζ-LEYP	1 014.7	[919.7, 1 109.7]
		NHPP	989.9	[928.3, 1 051.6]
		Observation	990	
2004 – 2006	84 421	ζ-LEYP	395.8	[355.3, 436.2]
		NHPP	388.8	[350.1, 427.4]
		Observation	411	

Tab. 9.14 – Comparaison des nombres de défaillances observés et prédits pour la fonte ductile avec les modèle ζ-LEYP et NHPP

9.5.3 Validation

Les prédictions de défaillances des canalisations en fonte ductile entre le 01/01/2004 et le 31/12/2006, calculées grâce au modèle calé sur les défaillances observées du 01/01/1995 au 31/12/2003, sont comparées aux défaillances réellement observées sur la même période aux fins de validation du pouvoir prédictif des modèles ζ-LEYP et NHPP. La courbe de performance

prédictive obtenue conformément à la méthode présentée en sous-section 5.5.1 est portée à la figure 9.2 ; la performance des deux modèles, comparativement à ce qui est observé pour la fonte grise, est sensiblement plus élevée.

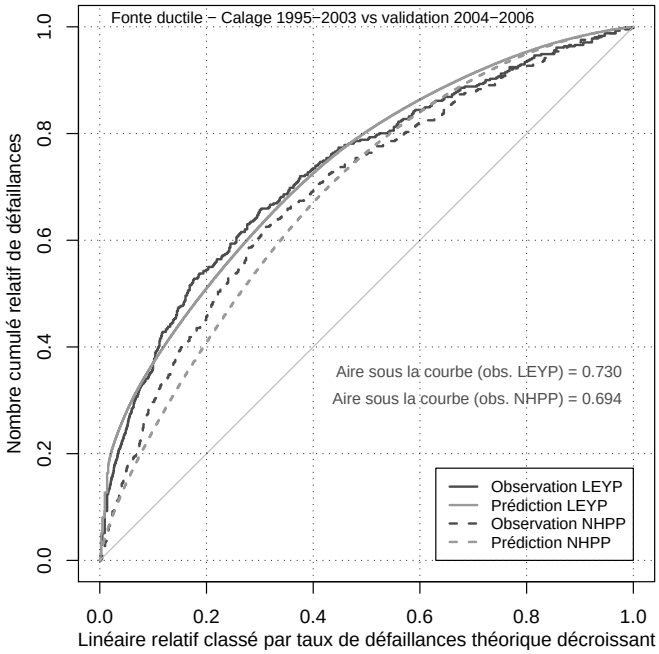


FIG. 9.2 – Performance prédictive du modèle fonte ductile

Le modèle ζ-LEYP permettrait d'orienter efficacement les renouvellements sur les tronçons les plus à risque, comme en témoigne le tableau 9.15 portant les proportions de défaillances qui auraient été évitées si la proportion correspondante du linéaire avait été renouvelée, en supposant le renouvellement prioritaire des tronçons que le modèle classe comme étant les plus à risque. Comme pour la fonte grise, la performance du NHPP apparaît nettement en retrait sur celle du ζ-LEYP.

% linéaire renouvelé		0.1	0.5	1.0	5.0
% défaillances évitables avec	ζ-LEYP	1.2	5.4	10.0	24.6
	NHPP	0.0	1.2	3.4	16.8

TAB. 9.15 – Proportions de défaillances évitables pour une proportion donnée de linéaire de fonte ductile renouvelée selon les modèles ζ-LEYP et NHPP

La figure 9.3 récapitule graphiquement, pour les 5 % du linéaire les plus à risque (respectivement au sens du ζ-LEYP et du NHPP), la supériorité de la performance prédictive du ζ-LEYP sur celle du NHPP.

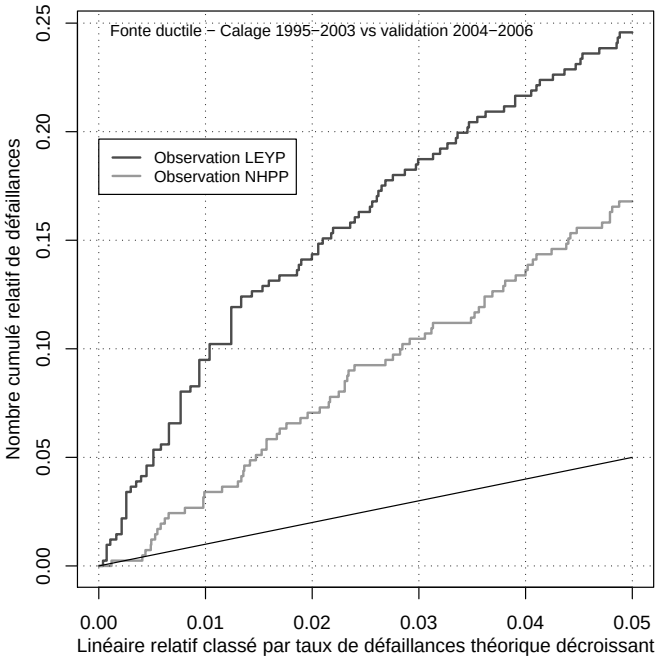


FIG. 9.3 – Performance prédictive du modèle fonte ductile pour les 5 % du linéaire les plus à risque

Chapitre 10

Conclusion et perspectives

La première partie de ce travail, consacrée à des développements théoriques, a montré comment l'extension linéaire du processus de Yule permet de construire un processus stochastique de comptage proche à bien des points de vue du NHPP, mais doté d'une mémoire des événements passés. Les distributions binomiales négatives, marginale et conditionnelle, de la fonction de comptage du LEYP ont été établies, offrant la possibilité d'effectuer des calculs de prédictions faciles à mettre en oeuvre. Une version du modèle LEYP, basée sur la fonction d'intensité dite de Yule-Weibull-Cox, a été particulièrement étudiée, dans l'optique de son application aux données de défaillances de canalisations d'eau. Ce modèle suppose que le risque instantané de défaillance augmente comme une puissance de l'âge des canalisations, et permet de prendre en compte l'effet proportionnel de facteurs explicatifs du risque (covariables). La fonction de vraisemblance des paramètres du modèle LEYP a été construite, et la possibilité d'estimer les paramètres du modèle en maximisant la vraisemblance a été étudiée. Afin de pouvoir traiter des données concernant des canalisations parfois très anciennes, dont les chances d'avoir été maintenues en service jusqu'à la fenêtre d'observation dépendent vraisemblablement de leur robustesse, une variante du modèle LEYP, dénommé ζ-LEYP, a été étudiée, en considérant une probabilité de maintien en service suite à une défaillance fonction de l'âge de la canalisation. La distribution de la fonction de comptage ainsi que la vraisemblance du modèle ζ-LEYP ont aussi été étudiées.

La seconde partie du travail a consisté à illustrer l'utilisation pratique du modèle ζ-LEYP sur deux jeux de données réelles, de grande taille, relatifs aux canalisations en fontes grise et ductile d'un important syndicat de distribution d'eau potable du nord de la France. Les résultats valident la possibilité de caler les paramètres du modèle en conditions opérationnelles, de mettre en évidence l'effet significatif de facteurs de risque techniquement pertinents, et montrent des performances prédictives du modèle supérieures au modèle NHPP pris comme référence ; le modèle ζ-LEYP s'avère ainsi être un outil performant pour bâtir un programme annuel de renouvellement de canalisations, ainsi que pour comparer, au moins sur le moyen terme, des stratégies de gestion patrimoniale.

Le modèle ζ-LEYP est dores et déjà utilisé dans le noyau de calcul d'un logiciel de prédiction statistique des défaillances de canalisations de distribution d'eau, commercialisé depuis fin 2007 auprès de gestionnaires de réseaux, de services publics ayant mission de régulation, et de bureaux d'études.

Des recherches complémentaires devront s'attacher à la sensibilité du modèle tant vis à vis

de ses paramètres structuraux :

- α, dans le cas où la valeur de ce paramètre tend vers 0 ;
- δ, dans le cas où la valeur de ce paramètre tend vers 1 ;
- ζ_0, dans le cas où la valeur de ce paramètre tend vers $-\infty$, ζ_1 étant fixé à 1 ;
- ζ_1, dans le cas où la valeur de ce paramètre tend vers 0 ;

que d'incertitudes quant à la valeur de deux caractéristiques essentielles des conduites :

- l'âge ;
- la longueur des tronçons.

Les perspectives d'amélioration de cet outil de modélisation concernent :

- le développement d'un test inférentiel de qualité d'ajustement du modèle ;
- la performance numérique de la méthode de maximisation de la vraisemblance (actuellement un algorithme de Levenberg-Marquardt) ;
- la prise en compte de facteurs de risque dépendant du temps, tels que les facteurs météorologiques ou des variations temporelles des pratiques de gestion du réseau (par exemple, les campagnes de recherches de fuites) ;
- la prise en compte d'éventuelles corrélations spatiales du risque de défaillance de conduites plus ou moins voisines.

Une voie possible pour développer un test inférentiel de qualité d'ajustement, dont la carence a été évoquée en section 5.4, pourrait consister en la mise au point d'un test de χ^2 de comparaison pour l'ensemble des tronçons d'un échantillon des distributions observée et théorique du nombre de défaillances par tronçon durant la fenêtre d'observation ; le matériel théorique approprié à cette recherche est exposé dans l'ouvrage de référence *A Guide to Chi-Squared Testing* de P.E. Greenwood et M.S. Nikulin [Greenwood et Nikulin, 1996].

Concernant la méthode de maximisation de la vraisemblance, il serait particulièrement intéressant d'évaluer la performance de la méthode de quasi-Newton dite « BFGS », publiée en 1970 par ses promoteurs Broyden, Fletcher, Goldfarb et Shanno, et bien décrite par [Nocedal et Wright, 1999]. Selon les premiers essais effectués avec la fonction `optim()` du module STATS de l'environnement R, documenté dans [R Development Core Team, 2006], la version dénommée `L-BFGS-B` publiée par [Byrd *et al.*, 1995], qui permet de définir des contraintes sur les paramètres, semble particulièrement attractive, simple d'emploi et sensiblement plus rapide que celle de Nelder-Mead, quoique plus compliquée à programmer en cas d'utilisation hors de l'environnement R.

La prise en compte de facteurs de risque dépendant du temps est une pratique courante dans l'analyse de données épidémiologiques ou bio-médicales. Le modèle le plus utilisé est le modèle semi-paramétrique de Cox ; des exemples sont rapportés, étudiés sous SAS ou sous R par [Therneau et Grambsch, 2000]. Il convient cependant de noter que ces études portent toujours sur des cas où la covariable dépendante du temps est une indicatrice permettant de distinguer les périodes où un patient est soumis à un traitement ou exposé à un facteur de risque. Le cas d'une covariable météorologique, comme par exemple la température de l'eau injectée dans le réseau de distribution, est qualitativement différent d'un double point de vue :

- la covariable est attachée au temps plus qu'aux individus qui subissent, et ce de façon simultanée, l'effet de la covariable ;
- la valeur de la covariable peut varier sur un pas de temps très court (le jour).

Les recherches menées actuellement au Cemagref de Bordeaux s'attachent à la mise au point d'un algorithme numériquement efficace pour calculer la fonction de vraisemblance exprimée

par la proposition 6.9 lorsque $\Lambda(t)$ implique d'intégrer sur un intervalle de temps long une fonction de données météorologiques constante par morceaux d'une journée ou d'une décade.

La prise en compte de corrélations spatiales du risque de défaillance de conduites plus ou moins éloignées les unes des autres est difficile du fait que la dimension topologique d'un réseau de canalisation est intermédiaire entre celle d'un objet linéaire et celle d'une surface, alors que les méthodes classiques de la géostatistique concernent des phénomènes en dimension 2, voire 3. La littérature consacrée à la distribution spatiale des défaillances d'un réseau d'eau est en outre peu abondante ; la voie proposée par [Goulter *et al.*, 1993], consistant à étudier la densité de défaillances dans un disque centré sur une défaillance observée initialement, en fonction du rayon, mériterait sans doute d'être prise en considération. L'ajout à la fonction d'intensité du LEYP d'un facteur de risque individuel, à l'instar des modèles dits « de fragilité », en définissant la distribution conjointe de ce facteur à l'aide d'une matrice des covariances entre les tronçons fonction de la matrice des distances entre tronçons mériterait d'être explorée ; une difficulté majeure réside cependant dans l'impossibilité de manipuler d'un seul bloc une matrice des distances complète pour un réseau de plusieurs milliers, voire dizaines de milliers de tronçons.

Appendices

Annexe A

Identité utilisée pour prouver la proposition 3.2

Proposition A.1.

$$-\sum_{k=0}^{m-1} \frac{1}{\prod_{l=0,l\neq k}^{m} (\alpha_l - \alpha_k)} = \frac{1}{\prod_{l=0,l\neq m}^{m} (\alpha_l - \alpha_m)} \tag{A.1}$$

Preuve :
Soit f une fonction quelconque, et $D_0 = 1$, $D_1 = x - \alpha_0$, $D_2 = (x - \alpha_0)(x - \alpha_1)$, ..., $D_m = (x - \alpha_0)(x - \alpha_1)\dots(x - \alpha_{m-1})$ les m premiers polynômes de Newton, la suite de réels $S = \{\alpha_j, j = 0, 1, \dots, m-1\}$ étant incluse dans un intervalle borné, et supposée croissante (sans perte de généralité pour notre propos).
Rappelons que le polynôme interpolateur de Lagrange de f sur S est défini par :

$$L_S(f) = \sum_{i=0}^{m} \Delta_i(\alpha_0, \dots, \alpha_i) D_i$$

où les différences divisées $\Delta_i(\alpha_0, \dots, \alpha_i)$ de f par rapport à S sont définies par récurrence :

$$\begin{aligned}
\Delta_0(\alpha_0) &= f(\alpha_0) \\
\Delta_1(\alpha_0, \alpha_1) &= \frac{f(\alpha_0) - f(\alpha_1)}{\alpha_0 - \alpha_1} \\
&\dots \\
\Delta_i(\alpha_0, \dots, \alpha_i) &= \frac{\Delta_{i-1}(\alpha_0, \dots, \alpha_{i-1}) - \Delta_{i-1}(\alpha_1, \dots, \alpha_i)}{\alpha_0 - \alpha_i}
\end{aligned}$$

Notre démonstration repose sur le résultat classique suivant, relatif aux différences divisées :

$$\Delta_i(\alpha_0, \dots, \alpha_i) = \sum_{k=0}^{i} \frac{f(\alpha_k)}{\prod_{l=0,l\neq k}^{i} (\alpha_k - \alpha_l)} \tag{A.2}$$

En effet, choisissons f telle que $f(\alpha_i) = 1,\ \forall i \in \{0, 1, \ldots, m-1\}$. L'équation (A.2) devient alors à l'ordre m :

$$\Delta_m(\alpha_0, \ldots, \alpha_m) = \sum_{k=0}^{m} \frac{1}{\prod_{l=0, l\neq k}^{m} (\alpha_k - \alpha_l)} = \sum_{k=0}^{m} \frac{(-1)^m}{\prod_{l=0, l\neq k}^{m} (\alpha_l - \alpha_k)}$$

Comme par ailleurs, mis à part à l'ordre 0, toutes les différences divisées sont alors nulles, il vient :

$$\sum_{k=0}^{m} \frac{1}{\prod_{l=0, l\neq k}^{m} (\alpha_l - \alpha_k)} = 0$$

ou de façon équivalente :

$$-\sum_{k=0}^{m-1} \frac{1}{\prod_{l=0, l\neq k}^{m} (\alpha_l - \alpha_k)} = \frac{1}{\prod_{l=0, l\neq m}^{m} (\alpha_l - \alpha_m)} \qquad \square$$

Annexe B

Gradient et Hessienne

La vraisemblance $L(\boldsymbol{\theta})$ d'un processus théorique LEYP de vecteur de paramètres $\boldsymbol{\theta}$ sachant qu'une séquence de m défaillances sont intervenues aux instants $t_1 < \ldots < t_j < \ldots < t_m$ dans l'intervalle d'observation $[a, b]$ est abrégée en L dans la suite, et $\psi(.)$ est la dérivée d'ordre un de la fonction log-gamma à l'instar de [Abramowitz et Stegun, 1972].

Aux fins d'estimation des paramètres α, δ et $\boldsymbol{\beta} = (\beta_k)$, nous avons besoin des dérivées partielles premières et secondes explicites de $\ln L$ par rapport à ces paramètres, dont les formes analytiques font l'objet de cette annexe :

$$\ln L = m \ln \alpha + \ln \Gamma\left(\frac{1}{\alpha} + m\right) - \ln \Gamma\left(\frac{1}{\alpha}\right) + m \ln \delta + m \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{\beta} + (\delta - 1) \sum_{j=1}^{m} \ln t_j + \alpha e^{\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{\beta}} \sum_{j=1}^{m} t_j^{\delta} - \left(\frac{1}{\alpha} + m\right) \ln C$$

$$\text{avec :} \quad A = \alpha a^{\delta} e^{\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{\beta}}, \quad B = \alpha b^{\delta} e^{\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{\beta}} \quad \text{et} \quad C = e^{B} - e^{A} + 1$$

$$\frac{\partial [\ln L]}{\partial \alpha} = \frac{m}{\alpha} - \frac{\psi(\alpha^{-1} + m)}{\alpha^2} + \frac{\psi(\alpha^{-1})}{\alpha^2} + e^{\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{\beta}} \sum_{j=1}^{m} t_j^{\delta} + \frac{\ln C}{\alpha^2} - \left(\frac{1}{\alpha} + m\right) \frac{1}{C} \frac{\partial C}{\partial \alpha}$$

$$\text{avec :} \quad \frac{\partial C}{\partial \alpha} = \frac{1}{\alpha} \left(B e^{B} - A e^{A} \right)$$

$$\frac{\partial [\ln L]}{\partial \delta} = \frac{m}{\delta} + \sum_{j=1}^{m} \ln t_j + \alpha e^{\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{\beta}} \sum_{j=1}^{m} t_j^{\delta} \ln t_j - \left(\frac{1}{\alpha} + m\right) \frac{1}{C} \frac{\partial C}{\partial \delta}$$

$$\text{avec :} \quad \frac{\partial C}{\partial \delta} = (\ln B) B e^{B} - (\ln A) A e^{A}$$

$$\frac{\partial [\ln L]}{\partial \beta_k} = m Z_k + \alpha Z_k e^{\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{\beta}} \sum_{j=1}^{m} t_j^{\delta} - \left(\frac{1}{\alpha} + m\right) \frac{1}{C} \frac{\partial C}{\partial \beta_k}$$

$$\text{avec :} \quad \frac{\partial C}{\partial \beta_k} = Z_k \left(B e^{B} - A e^{A} \right)$$

$$\frac{\partial^2[\ln L]}{\partial\alpha^2} = -\frac{m}{\alpha^2} + \frac{\psi'(\alpha^{-1}+m)}{\alpha^4} + 2\frac{\psi(\alpha^{-1}+m)}{\alpha^3} - \frac{\psi'(\alpha^{-1})}{\alpha^4} - 2\frac{\psi(\alpha^{-1})}{\alpha^3} + \frac{2}{C\alpha^2}\frac{\partial C}{\partial\alpha} - \frac{2\ln C}{\alpha^3} + \left(\frac{1}{\alpha}+m\right)\left(\frac{1}{C^2}\left(\frac{\partial C}{\partial\alpha}\right)^2 - \frac{1}{C}\frac{\partial^2 C}{\partial\alpha^2}\right)$$

$$\text{avec :} \quad \frac{\partial^2 C}{\partial\alpha^2} = \frac{1}{\alpha}\left(B^2 e^B - A^2 e^A\right)$$

$$\frac{\partial^2[\ln L]}{\partial\alpha\partial\delta} = e^{\mathbf{z}^\top\boldsymbol{\beta}}\sum_{j=1}^{m} t_j^\delta \ln t_j + \frac{1}{C\alpha^2}\frac{\partial C}{\partial\delta} + \left(\frac{1}{\alpha}+m\right)\left(\frac{1}{C^2}\frac{\partial C}{\partial\alpha}\frac{\partial C}{\partial\delta} - \frac{1}{C}\frac{\partial^2 C}{\partial\alpha\partial\delta}\right)$$

$$\text{avec :} \quad \frac{\partial^2 C}{\partial\alpha\partial\delta} = \frac{1}{\alpha}\left((1+B)(\ln B)Be^B - (1+A)(\ln A)Ae^A\right)$$

$$\frac{\partial^2[\ln L]}{\partial\alpha\partial\beta_k} = Z_k e^{\mathbf{z}^\top\boldsymbol{\beta}}\sum_{j=1}^{m} t_j^\delta + \frac{1}{C\alpha^2}\frac{\partial C}{\partial\beta_k} + \left(\frac{1}{\alpha}+m\right)\left(\frac{1}{C^2}\frac{\partial C}{\partial\alpha}\frac{\partial C}{\partial\beta_k} - \frac{1}{C}\frac{\partial^2 C}{\partial\alpha\partial\beta_k}\right)$$

$$\text{avec :} \quad \frac{\partial^2 C}{\partial\alpha\partial\beta_k} = \frac{Z_k}{\alpha}\left((1+B)Be^B - (1+A)Ae^A\right)$$

$$\frac{\partial^2[\ln L]}{\partial\delta^2} = -\frac{m}{\delta^2} + \alpha e^{\mathbf{z}^\top\boldsymbol{\beta}}\sum_{j=1}^{m} t_j^\delta(\ln t_j)^2 + \left(\frac{1}{\alpha}+m\right)\left(\frac{1}{C^2}\left(\frac{\partial C}{\partial\delta}\right)^2 - \frac{1}{C}\frac{\partial^2 C}{\partial\delta^2}\right)$$

$$\text{avec :} \quad \frac{\partial^2 C}{\partial\delta^2} = (1+B)(\ln B)^2 Be^B - (1+A)(\ln A)^2 Ae^A$$

$$\frac{\partial^2[\ln L]}{\partial\delta\partial\beta_k} = \alpha Z_k e^{\mathbf{z}^\top\boldsymbol{\beta}}\sum_{j=1}^{m} t_j^\delta \ln t_j + \left(\frac{1}{\alpha}+m\right)\left(\frac{1}{C^2}\frac{\partial C}{\partial\delta}\frac{\partial C}{\partial\beta_k} - \frac{1}{C}\frac{\partial^2 C}{\partial\delta\partial\beta_k}\right)$$

$$\text{avec :} \quad \frac{\partial^2 C}{\partial\delta\partial\beta_k} = Z_k\left((1+B)(\ln B)Be^B - (1+A)(\ln A)Ae^A\right)$$

$$\frac{\partial^2[\ln L]}{\partial\beta_k\partial\beta_l} = \alpha Z_k Z_l e^{\mathbf{z}^\top\boldsymbol{\beta}}\sum_{j=1}^{m} t_j^\delta + \left(\frac{1}{\alpha}+m\right)\left(\frac{1}{C^2}\frac{\partial C}{\partial\beta_k}\frac{\partial C}{\partial\beta_l} - \frac{1}{C}\frac{\partial^2 C}{\partial\beta_k\partial\beta_l}\right)$$

$$\text{avec :} \quad \frac{\partial^2 C}{\partial\beta_k\partial\beta_l} = Z_k Z_l\left((1+B)Be^B - (1+A)Ae^A\right)$$

Bibliographie

M. Abramowitz et I. Stegun : *Handbook of mathematical functions*. Dover Publications, Inc., New York, 12th édition, 1972.

P. K. Andersen, Ø. Borgan, R. D. Gill et N. Keiding : *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1st édition, 1993.

Y. Bard : *Nonlinear Parameter Estimation*. Academic Press, New York, 1974.

A. T. Bharucha-Reid : *Elements of the Theory of Markov Processes and Their Applications*. McGraw-Hill Book Company, New York, 1960.

A. T. Bharucha-Reid : *Elements of the Theory of Markov Processes and Their Applications*. Dover Publications, Inc., Mineola, New York, 1997.

R.H. Byrd, P. Lu, P. Nocedal et C. Zhu : A limited memory algorithm for bound constraints optimization. *SIAM Journal of Scientific Computing*, 16:1190–1208, 1995.

C. C. H. Chang, W. Chan et A. S. Kapadia : The analysis of recurrent failure times : The time-dependent yule process approach. *Reliability and Survival Analysis. Commun. Statist. - Theory Meth.*, 31(7):1203–1213, 2002.

R. J. Cook et J. F. Lawless : Marginal analysis of recurrent events and a terminating event. *Statistics in Medecine*, 16:911–924, 1997.

R. J. Cook et J. F. Lawless : Analysis of repeated events. *Statistical Methods in Medical Research*, 11:141–166, 2002.

D. R. Cox : *Renewal Theory*. Chapman and Hall, New York, 1962.

D. R. Cox : Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

D. R. Cox et D. V. Hinkley : *Theoretical Statistics*. Chapman and Hall, London, 1974.

A. Debón, A. Carrión, E. Cabrera et H. Solano : Comparing risk of failure models in water supply networks using roc curves. *Reliability Engineering and System Safety*, 95:43–48, 2010.

P. Eisenbeis : *Modélisation statistique de la prévision des défaillances sur les conduites d'eau potable*. Thèse de doctorat, Université Louis Pasteur Strasbourg, 1994.

R. Fletcher : *Practical Methods of Optimisation*. John Wiley and Sons, Chichester, 2 édition, 1987.

I. Goulter, J. Davidson et P. Jacobs : Predicting water-main breakage rate. *Journal of Water Resources Planning and Management*, 119(4):419–436, 1993.

M. Greenwood et G. U. Yule : An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society, Series A*, 83:255–279, 1920.

P. E. Greenwood et M. S. Nikulin : *A Guide to Chi-Squared Testing*. John Wiley and Sons, New York, 1996.

Y. Kleiner et B. Rajani : Comprehensive review of structural deterioration of water mains : Statistical models. *Urban Water*, 3(3):131–150, 2001.

J. F. Lawless : Regression methods for poisson process data. *Journal of the American Statistical Association, Theory and Methods*, 82(399):808–815, 1987.

Y. Le Gat : Forecasting pipe failures in drinking water networks using stochastic processes models - respective relevance of renewal and poisson processes. *In Proceedings of the 13th European Junior Scientist Workshop*, pages 105–117, Dresden, Germany, septembre 8-12 1999. TU Dresden.

Y. Le Gat : Evaluation de la performance d'un modèle de prévision des casses en réseau d'adduction d'eau potable. Mémoire de DEA Epidémiologie et Intervention en Santé Publique, Université de Bordeaux 2, 2002.

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

P. Le Gauffre, J. Ruffier, C. Tanguy, K. Laffrechine, M. Miramond, L. Perraudin et L. Richard : Projet CAPTUR - consolidation d'un cadre théorique d'analyse des patrimoines techniques urbains de type réseau. Rapport final de recherche. Ministère de l'Education nationale, de la Recherche et de la Technologie - Action concertée incitative Ville, décision d'aide 99v0492, INSA, Lyon, France, 2001.

J. Nocedal et S.J. Wright : *Numerical Optimization*. Springer, 1999.

G. Pelletier : *Impact du remplacement des conduites d'acqueduc sur le nombre annuel de bris*. Thèse de doctorat, Université du Québec, INRS-Eau, 1999.

W. H. Press, S. A. Teukolsky, W. T. Vetterling et B. P. Flannery : *Numerical Recipes in C - The Art of Scientific Computing*. Cambridge University Press, New York, USA, 2 édition, 2002.

R Development Core Team : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http ://www.R-project.org`. ISBN 3-900051-07-0.

C. R. Rao : *Linear Statistical Inference and its Applications*. John Wiley, New York, 2nd édition, 1973.

A. Renyi : *Calcul des probabilités*. Dunod, Paris, 1966.

S. Ross : *Stochastic Processes*. John Wiley and Sons, New York, 1983.

S. Ross : *Simulation*. Academic Press, San Diego, 2nd édition, 1997.

J. Røstum : Statistical modelling of pipe failures in water networks. Doctor engineer dissertation, Norwegian University of Science and Technology, Department of Hydraulic and Environmental Engineering, Trondheim, Norway, 2000.

O. Samset : *Reliability Estimation Based on Operating History of Repairable Systems*. Thèse de doctorat, The Norwegian Institute of Technology, Division of Mathematical Sciences, 1994.

SAS Institute Inc. : *SAS OnlineDoc®, Version 8*. SAS Institute Inc., Cary, NC, USA, 1999.

A. Sen et N. Balakrishnan : Convolution of geometrics and a reliability problem. *Statistics and Probability Letters*, 43:421–426, 1999.

T. M. Therneau et P. M. Grambsch : *Modeling Survival Data : Extending the Cox Model*. Springer-Verlag, New York, 2000.

S. Yamijala, S. D. Guikema et K. Brumbelow : Statistical models for the analysis of water distribution system pipe break data. *Reliability Engineering and System Safety*, 94:282–293, 2009.

Index