Analyse et modélisation statistique de données de consommation électrique

Kévin Jaunâtre

THESE DE DOCTORAT DE

L'UNIVERSITE BRETAGNE SUD
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Mathématiques et leurs Interactions

Par
Kévin Jaunâtre

Analyse et modélisation statistique de données de consommation électrique

Thèse présentée et soutenue à Vannes, le 18/01/2019
Unité de recherche : UMR CNRS 6205
Thèse N° : 520

Rapporteurs avant soutenance :

Christian Derquenne	Chargé de recherche	EDF R&D
Philippe Naveau	Directeur de recherche	LSCE
Denys Pommeret	Professeur	I2M

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition ne comprend que les membres présents

Pierre Ailliot Examinateur	Maître de Conférences	LMBA
Christian Derquenne Rapporteur	Chargé de recherche	EDF R&D
Gilles Durrieu Directeur de thèse	Professeur	LMBA
Ion Grama Directeur de thèse	Professeur	LMBA
Valérie Monbet Examinateur	Professeur	IRMAR
Philippe Naveau Rapporteur	Directeur de recherche	LSCE
Denys Pommeret Rapporteur	Professeur	I2M
Pierre Ribereau Examinateur	Maître de Conférences	ICJ

Analyse et modélisation statistique de données de consommation électrique

Résumé :

En octobre 2014, l'Agence De l'Environnement et de la Maîtrise de l'Energie (ADEME) en coopération avec l'entreprise ENEDIS (anciennement ERDF pour Électricité Réseau Distribution France) a démarré un projet de recherche dénommé "smart-grid SOLidarité-ENergie-iNovation" (SOLENN) avec comme objectifs l'étude de la maîtrise de la consommation électrique par un accompagnement des foyers et la sécurisation de l'approvisionnement électrique entre autres. Cette thèse s'inscrit dans le cadre des objectifs susnommés. Le projet SOLENN est piloté par l'ADEME et s'est déroulé sur la commune de Lorient. Le projet a pour but de mettre en œuvre une pédagogie pour sensibiliser les foyers aux économies d'énergie.

Dans ce contexte, nous abordons une méthode d'estimation des quantiles extrêmes et des probabilités d'événements rares pour des données fonctionnelles non-paramétriques qui fait l'objet d'un package R. Nous proposons ensuite une extension du fameux modèle de Cox à hasards proportionnels et permet l'estimation des probabilités d'événements rares et des quantiles extrêmes.

Enfin, nous donnons l'application de certains modèles statistique développés dans ce document sur les données de consommation électrique et qui se sont avérés utiles pour le projet SOLENN. Une première application est en liaison avec le programme d'écrêtement mené par ENEDIS afin de sécuriser le fonctionnement du réseau électrique. Les foyers sont sujets à une modulation ou une diminution de la puissance appelée pendant une période donnée. Le but est d'étudier le comportement des utilisateurs pendant cette période. Une deuxième application est la mise en place du modèle linéaire pour étudier l'effet de plusieurs visites individuelles sur la consommation électrique. Le but est de chercher un impact sur la consommation électrique des individus pour la semaine (le mois) suivant une visite.

Mots clés : Théorie des valeurs extrêmes, Analyse de survie, Queues lourdes, Censure à droite, Modèle de Cox.

Electric consumption data modeling and analysis

Abstract :

In October 2014, the French Environment & Energy Management Agency with the ENEDIS company started a research project named SOLENN ("SOLidarité ENergie iNovation") with multiple objectives such as the study of the control of the electric consumption by following the households and to secure the electric supply. The SOLENN project was lead by the ADEME and took place in Lorient, France. The main goal of this project is to improve the knowledge of the households concerning the saving of electric energy.

In this context, we describe a method to estimate extreme quantiles and probabilites of rare events which is implemented in a R package. Then, we propose an extension of the famous Cox's proportional hazards model which allows the etimation of the probabilites of rare events.

Finally, we give an application of some statistics models developped in this document on electric consumption data sets which were useful for the SOLENN project. A first application is linked to the electric constraint program directed by ENEDIS in order to secure the electric network. The houses are under a reduction of their maximal power for a short period of time. The goal is to study how the household behaves during this period of time. A second application concern the utilisation of the multiple regression model to study the effect of individuals visits on the electric consumption. The goal is to study the impact on the electric consumption for the week or the month following a visit.

Key words : Extreme value theory, Survival analysis, Heavy tails, Right censoring, Cox's model.

Remerciements

Ce paragraphe me permet d'exprimer ma gratitude envers toutes les personnes qui m'ont soutenu lors de ces trois années et ont rendu cette thèse possible.

Je remercie tout particulièrement mes deux directeurs de thèse, M. Gilles Durrieu et M. Ion Grama pour la chance que vous m'avez donnée de travailler à vos côtés. Vous m'avez fait partager votre enthousiasme et votre passion pour le domaine de la recherche que je ne connaissais pas. Je vous remercie également des conseils que vous avez pu me prodiguer et du temps que vous m'avez consacré malgré vos nombreuses responsabilités.

Je souhaite également remercier les rapporteurs de cette thèse M. Christian Derquenne, M. Philippe Naveau et M. Denys Pommeret d'avoir accepté de relire ce manuscrit et pour leurs commentaires constructifs. Je remercie les membres du jury M. Pierre Ailliot, Mme Valérie Monbet et Pierre Ribereau d'avoir accepté de prendre le temps d'évaluer mon travail.

Je remercie l'ADEME et Enedis qui ont permis à cette thèse de voir le jour grâce au projet SOLENN. Je remercie l'ensemble des partenaires du projet et plus particulièrement ceux avec qui j'ai eu un échange quasi-quotidiens pendant ces trois années. Un grand merci à Claudie, Florent, Élise, Hervé, Guillaume, Morgane et Julie.

Merci également aux membres du LMBA pour leur bonne humeur. Je remercie mes compagnons doctorants : À Éric et nos conversations sportives, à Jonathan et ses conseils très productifs, à Ronan et sa positivité journalière et nos sorties sur Vannes, à Phu pour nos cafés réguliers, à Du pour sa cuisine vietnamienne délicieuse, à Jamila pour avoir pris plaisir à modifier ma playlist, à Salim pour nos parties d'échec endiablées, à Tarik, Cuong, Thong, Thuy, Thao, Xiao, Erwan et Hélène. Merci également aux autres membres du laboratoire et plus particulièrement Véronique ainsi que Sandrine pour m'avoir énormément aidé avec les tâches administratives pour lesquels vous savez mon attachement.

Merci à tous mes amis proches pour leur soutien. Je remercie également mon club de tennis de table ainsi que TySquash.

Enfin, je souhaite remercier ma famille pour son support tout au long de la thèse. À mes grands-mères, mes oncles, tantes, cousins et cousines. À ma petite soeur et à mes parents pour avoir cru en moi.

Table des matières

Glossaire

AC Foyers considérés avec un mode de chauffage principal non électrique. 78, 80, 84, 85, 87–90, 92, 93, 95–101

ADEME Agence De l'Environnement et de la Maîtrise de l'Energie. XII

ALOEN Agence LOcale de l'ENergie. 102

ANAH Agence NAtionale de l'Habitat. XII

C Foyers appartenant au panel collectif. 80, 87, 89, 91, 92

CE Foyers considérés avec un mode de chauffage principal électrique. 78, 80, 84, 86–88, 90, 91, 95–101

E Foyers appartenant au panel témoin. 80, 87, 91

IA Foyers appartenant au panel individuel ALOEN. 80, 87, 91

ID Foyers appartenant au panel individuel Delta Dore. 80, 87, 89, 91, 92

IV Foyers appartenant au panel individuel Vity. 80, 85, 89, 92

PAM correspond à la Puissance Appelée Moyenne sur 10 minutes. Ce sont les données de courbes de charge. 14, 103

RTE Réseau de Transport d'électricité. XII

SOLENN SOLidarité-ENergie-iNovation. II, III, XII–XIV, 17, 113

Publications et Conférences

Articles :

- Gilles Durrieu, Ion Grama, Kévin Jaunâtre, Quang-Khoai Pham, and Jean-Marie Tricot. "extremefit : An R Package for Extreme Quantiles." *Journal of Statistical Software* (2016).
- Ion Grama and Kévin Jaunâtre. "Estimation of Extreme Survival Probabilities with Cox Model." arXiv preprint arXiv :1805.01638 (2018). En attente d'une deuxième réponse de la part des reviewers, *Statistics*.

Package R :

- extremefit : Estimation of Extreme Conditional Quantiles and Probabilities https://CRAN.R-project.org/package=extremefit

Conférences :

- extremefit : un package pour estimer les probabilités et quantiles conditionnels extrêmes -Gilles Durrieu, Ion Grama, Kévin Jaunâtre, Quang-Khoai Pham, et Jean-Marie Tricot, Laboratoire de Mathématiques de Bretagne Atlantique, Rencontres R 2016, Toulouse, 22-24 juin 2016.
- Étude des données de consommation électrique collectées avec les compteurs intelligents Linky - Gilles Durrieu, Ion Grama, Kévin Jaunâtre, et Jean-Marie Tricot, LMBA, Mathematics and Enterprises Days, Vannes, 12-13 avril 2018
- Extreme value estimation with Cox model - Ion Grama and Kevin Jaunatre, LMBA, Survival Analysis for Junior Researchers , Leiden, 24-26 avril 2018

Rapports du projet SOLENN :

- Livrable 8.4.2 : Restitution du traitement économétrique après le 1er hiver - Gilles Durrieu, Ion Grama, Kévin Jaunâtre, et Jean-Marie Tricot, ADEME, 23 mai 2017
- Livrable 8.4.3 : Restitution du traitement économétrique après le 2ème hiver - Gilles Durrieu, Ion Grama, Kévin Jaunâtre, et Jean-Marie Tricot, ADEME, 27 octobre 2017
- Livrable 8.4.3bis : Restitution du traitement économétrique après le 3ème hiver - Gilles Durrieu, Ion Grama, Kévin Jaunâtre, et Jean-Marie Tricot, ADEME, 10 juillet 2018

Présentation pour le projet SOLENN :

- 1er octobre 2015
- 9 février 2016
- 6 janvier 2017
- 6 juin 2017
- 26 octobre 2017
- 2 juin 2018
- 12 juillet 2018
- 14 septembre 2018

Introduction

Nous présentons dans ce chapitre le contexte de la thèse ainsi que les objectifs de celle-ci. Une brève présentation des travaux effectués jusqu'à présent ainsi qu'un résumé du projet associé à la thèse sont donnés. La structure de la thèse est fournie en fin de chapitre.

Contexte de la thèse

En 2010, un pacte électrique breton a été signé entre l'état, l'ADEME, RTE et l'ANAH. Ce pacte vise à proposer des solutions autour de trois actions complémentaires :

- des efforts importants de maîtrise de la demande en électricité.
- un développement ambitieux de la production d'énergie renouvelables.
- la sécurisation indispensable de l'alimentation électrique (production et réseaux).

Le projet SOLENN s'inscrit dans deux des trois volets du pacte électrique breton.

Le projet smart-grid SOLENN

Le projet smart-grid SOLENN (SOLidarité-ENergie-iNovation) est un projet accompagné par l'ADEME dans le cadre du programme Réseaux électriques intelligents des Investissements d'Avenir. Il est localisé sur l'agglomération de Lorient (France) et a débuté en octobre 2014 pour une période initiale de 3 ans. Douze partenaires [1] sont impliqués dont Enedis (anciennement ERDF), qui est le coordinateur.

[1] Enedis et RTE pour la filière électrique ; ALOEN, Lorient Agglomération, PEBreizh et Région Bretage pour les pouvoirs publics et les acteurs de la transition énergétique ; UFC que choisir et La CSF pour la représentation des consommateurs ; Delta Dore, Vity et Niji, qui sont des acteurs privés intervenant pour proposer des solutions domotiques et numériques qui sont testées dans le

Il s'inscrit dans la suite du déploiement des compteurs électriques communicants de nouvelle génération : les compteurs Linky.

Il s'agit d'un compteur communicant, cela signifie qu'il peut envoyer des données ou recevoir des ordres sans intervention physique. La pose des compteurs a commencé le 1^{er} décembre 2015 et l'objectif est de remplacer la totalité des anciens compteurs en France d'ici 2021. La communication du compteur fonctionne par CPL (courants porteurs en ligne). Le signal circule à travers les câbles du réseau électrique basse-tension jusqu'au poste de distribution du quartier où se trouve un concentrateur qui envoie les données au système d'information centralisé du distributeur basse-tension. Les données sont collectées une fois par jour, celles-ci sont détaillées au chapitre 1.4.1. Les compteurs Linky ont été déployés sur 10000 foyers de la commune de Lorient durant l'année 2015.

Le projet SOLENN se compose de deux volets parmi les trois du pacte électrique breton.

- **Maîtrise de la consommation d'électricité**. L'objectif est de mettre en œuvre une pédagogie pour sensibiliser les foyers aux économies d'énergie grâce à un accompagnement individuel et à des animations collectives.
- **Sécurisation de l'approvisionnement électrique**. Pour répondre à ce volet, le démonstrateur SOLENN s'appuie sur le développement et la mise en test d'une fonction de réduction de la puissance de coupure (écrêtement ciblé) sur une partie des foyers. Les objectifs sont la construction d'un processus permettant de réaliser l'écrêtement sur un groupe de foyers et de mesurer l'impact de cette nouvelle fonctionnalité en terme d'acceptabilité des particuliers.

Le projet a débuté en octobre 2014 pour une durée initiale de trois ans. À cause de certains retards, il a été décidé de prolonger le projet sur une année supplémentaire pour se terminer en septembre 2018. Avec cette prolongation, une deuxième vague de recrutement a été effectuée pendant l'hiver 2016/2017 pour avoir un échantillon plus conséquent.

Écrêtements

L'écrêtement ciblé est une modulation de la puissance appelée par le foyer pendant une période donnée. Il s'agit d'une diminution temporaire de la puissance maximale

projet ; et l'Université de Bretagne Sud associée à l'Université de Bretagne Occidentale et IMT Atlantique par des conventions de collaboration, en charge de l'analyse des données recueillies lors du programme.

électrique disponible dans le logement. En cas d'incident sur le réseau de distribution, et si l'appel aux différentes sources de production électrique ne permet pas d'assurer la demande du moment, la puissance électrique maximale souscrite par le foyer est temporairement baissée. Par exemple, un foyer avec un contrat d'une puissance maximale de 9 kVA pourrait voir celle-ci diminuer à 4.5 kVA pendant une période de quelques heures. L'écrêtement est testé en tant qu'alternative à l'effacement (coupure totale de l'électricité dans un foyer) qui est actuellement la seule possibilité en cas d'incident sur le réseau. Celui-ci permettrait, en période de crise, d'avoir une puissance abaissée sur 1000 foyers au lieu d'en avoir 300 sans électricité et 700 avec. Les foyers sont avertis par un sms de la baisse de puissance qui va arriver et de la durée de celle-ci. Le retour à la normale est également signalé.

Précédents travaux

Cette thèse vient s'ajouter aux travaux déjà réalisés dans plusieurs domaines. Différentes thèses ont été effectuées dans le cadre de la modélisation de consommation électriques. La thèse de Allab [1] porte sur la détection de ruptures dans la courbe de charge en modélisant celle-ci par un modèle de série temporelle de type ARMA. Celle de Sanquer [2] se focalise sur l'identification des appareils électro-domestiques à partir des signaux transitoires qu'ils génèrent ainsi que comment les caractériser et les identifier.

Pour la partie concernant les valeurs extrêmes, cette thèse suit le travail de la thèse de Quang Khoai Pham [3] qui propose une estimation des quantiles extrêmes et des probabilités d'événements rares conditionnellement à une covariable temporelle. Une étude théorique est menée dans la thèse de Ndao [4] sur l'estimation des quantiles extrêmes conditionnels en présence de censure et notamment en utilisant l'estimateur de Kaplan-Meier.

La problématique

Cette thèse s'inscrit dans le cadre du projet SOLENN et a deux objectifs :

- Étudier l'impact des visites sur les consommations électriques.
- Mettre en place un modèle permettant de détecter des foyers "à risque" pendant l'écrêtement ciblé.

Le premier objectif permet de mesurer l'efficacité des méthodes de maîtrise de la consommation électrique mises en place pendant le projet sur l'éventuelle modification de la consommation électrique. Le deuxième objectif a pour but d'avoir une

connaissance plus grande sur les écrêtements et de mettre en relation le pourcentage d'écrêtement avec le nombre de foyers qui pourraient subir une ouverture de breaker. Il est important qu'au moment de l'écrêtement, le nombre de foyers pouvant être impactés soit connu.

Pour répondre au second objectif, nous proposons une estimation des probabilités de survie en ajustant au modèle de Cox une méthode pour estimer les quantiles et probabilités de survie extrêmes.

Structure de la thèse

La présentation de ce travail s'articule en quatre chapitres. Dans un premier chapitre, nous présentons quelques notions sur la théorie des valeurs extrêmes ainsi que sur le modèle de Cox. Nous décrivons également les données sur lesquelles les résultats ont été construits. Dans un second chapitre, nous présentons un package R nommé **extremefit**. Celui-ci permet d'estimer les quantiles extrêmes conditionnellement à une variable temporelle. Dans un troisième chapitre, nous proposons un modèle pour estimer les probabilités de survie extrêmes en présence de covariables et de censures en utilisant le modèle de Cox. Dans un dernier chapitre, nous donnons les résultats des méthodes appliquées aux données de consommation électrique.

Chapitre 1

Rappels et description des données

Ce chapitre vise à rappeler quelques notions nécessaires à la bonne compréhension de la thèse. Nous présentons brièvement, dans un premier temps, une partie de la théorie des valeurs extrêmes sur laquelle nous nous appuyons. Pour une lecture plus approfondie sur cette théorie, nous référons à Beirlant et al. [5]. Nous rappelons ensuite le modèle bien connu de Cox [6] ainsi que quelques estimateurs, notamment celui de Nelson-Aalen (Nelson [7] et Aalen [8]) et de Kaplan-Meier [9]. De plus, nous présentons les données modélisées dans le chapitre 4.

1.1 Rappels sur la théorie des valeurs extrêmes

La théorie des valeurs extrêmes, communément appelée EVT (Extreme Value Theory), est une théorie utilisée dans le but d'étudier des événements qui apparaissent très rarement. Cela peut être utile dans de multiples domaines pratiques tels que les intempéries, nous pouvons citer entre autres Buishand et al. [10] ainsi que Gardes et Girard [11], la finance avec plusieurs applications dans Danielsson et de Vries [12], McNeil [13], Embrechts et al. [14] ainsi que Gencay et Selcuk [15] ou l'assurance avec McNeil [16] et Rootzén et Tajvidi [17].

Supposons un échantillon $X_1, \cdots, X_n$ de n variables i.i.d. de distribution F. Le théorème central limite s'intéresse à la somme $S_n = X_1 + X_2 + ... + X_n$ et essaye de trouver des constantes $a_n > 0$ et b_n de telle sorte que $Y_n = a_n^{-1}(S_n - b_n)$ converge vers une distribution non dégénérée. Une fois la distribution limite connue, elle est utilisée pour approcher la quantité Y_n. Généralement, la limite de cette somme S_n est une distribution normale. Si la distribution de $X_1, ..., X_n$ est une distribution à queue lourde, alors la limite est une autre distribution. Par exemple, si F est une distribution de type Pareto, alors la moyenne ne va pas avoir une distribution limite normale. Les valeurs extrêmes produites par les distributions de type Pareto vont

affecter la moyenne de telle sorte qu'un comportement différent de la distribution normale est obtenu.

1.1.1 Les limites possibles

Considérons le maximum

$$X_{(n)} = \max\{X_1, X_2, ..., X_n\}.$$

Là où le TCL établit que l'espérance de la variable aléatoire X converge vers une distribution normale indépendamment de la loi de X (lorsque les moments d'ordre 1 et 2 existent), la théorie des valeurs extrêmes donne des résultats sur le comportement asymptotique de $X_{(n)}$ et des excès au-delà d'un seuil τ. Il n'est pas nécessaire de connaître la distribution F pour avoir les deux résultats suivants. Le théorème suivant donne la loi limite du maximum.

Theorem 1.1.1. *(Fisher-Tippett-Gnedenko-Von Mises-Jenkinson 1927)*

Sous certaines conditions de régularités sur la fonction F, il existe un paramètre $\gamma \in \mathbb{R}$ et deux suites $(a_n)_{n\geq 1}$ positive et $(b_n)_{n\geq 1}$ réelle telles que pour tout $y \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{X_{(n)} - b_n}{a_n} \leq y\right] = \mathcal{H}_\gamma(y),$$

avec

$$\mathcal{H}_\gamma(y) = \begin{cases} e^{-(1+\gamma y)^{-1/\gamma}} \mathbb{1}_{\{1+\gamma y>0\}} & si \quad \gamma \neq 0, \\ e^{-e^{-y}} & sinon. \end{cases}$$

où γ est l'indice des valeurs extrêmes et $\mathcal{H}_\gamma$ est la loi des valeurs extrêmes.

Les suites $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ sont appelées des suites de normalisation. Dans Embrechts et al. [14], les auteurs s'intéressent à l'estimation de ces constantes pour plusieurs distributions suivies par la variable d'intérêt. Pour la distribution normale, les suites sont définies par $a_n = (2\ln n)^{-1/2}$ et $b_n = (2\ln n)^{1/2} - \frac{\ln\ln n + \ln 4\pi}{2(2\ln n)^{1/2}} + o\left((\ln n)^{-1/2}\right)$. Le théorème concernant la loi des excès au-delà d'un seuil est similaire au théorème précédent.

Theorem 1.1.2. *(Balkema-de Haan-Pickands 1974)*

S'il existe des constantes $\{a_n\}_{n\geq 0}$ positive et $\{b_n\}_{n\geq 0}$ réelle telles que

$$\mathbb{P}\left[\frac{X_{(n)} - b_n}{a_n} \leq y\right] \to \mathbb{G}(y) \quad quand \quad n \to \infty,$$

où $\mathbb{G}$ a la même forme que $\mathcal{H}_\gamma$. Alors,

$$\mathbb{P}\left[Y \leq y \,|\, Y > \tau\right] \to \mathbb{H}(y), \quad quand \quad \tau \to \tau_f,$$

où $\tau_f = \sup\{\tau \in \mathbb{R} : \mathbb{P}[Y \leq \tau] < 1\}$ *et*

$$\mathbb{H}(y) = 1 - \left[1 + \epsilon\left(\frac{y-\tau}{\sigma}\right)\right]_+^{-1/\gamma},$$

où $z_+ = \max(0, z)$ *et* $\sigma > 0$. *Les paramètres* σ *et* γ *correspondent respectivement aux paramètres d'échelle et de forme. Le paramètre* τ *est appelé seuil.*

La modélisation des valeurs extrêmes consiste à exploiter les deux théorèmes précédents pour avoir l'information sur la queue de la distribution de X. Ces deux théorèmes admettent le paramètre γ qui nous informe sur la forme de la fonction de répartition. Selon la valeur de γ, nous pouvons classer les lois en trois familles appelés **domaines d'attraction** :

- Si $\gamma < 0$, F appartient au domaine d'attraction de **Weibull**. Ce domaine contient les distributions dont le point terminal $x_f = \inf\{x, F(x) \geq 1\}$ est fini. Ces distributions peuvent être utilisées pour décrire la résistance d'un matériau et en fiabilité.
- Si $\gamma = 0$, F appartient au domaine d'attraction de **Gumbel**. Ce domaine regroupe les distributions dont la fonction de survie décroit vers zéro de façon exponentielle. Ces distributions sont souvent utilisées dans un contexte environnemental tel que l'hydrologie ou l'étude des tremblements de terre.
- Et si $\gamma > 0$, F appartient au domaine d'attraction de la loi de **Fréchet**. Les distributions dont la fonction de survie décroit comme une fonction polynomiale appartiennent à ce domaine. Ces distributions sont souvent utilisées en finance, en assurance ainsi qu'en météorologie pour l'étude de risque.

Nous donnons dans la Table 1.1 quelques exemples de distributions appartenant à chaque domaine d'attraction.

Fréchet-Pareto $\gamma > 0$	Gumbel $\gamma = 0$	Weibull $\gamma < 0$
Pareto	Weibull	Uniforme
GPD	Normale	Beta
Burr	Exponentielle	Reversed-Burr
Fréchet	Gamma	
Inverse-Gamma	Logistique	
Log-Gamma	Log-Normale	
Log-Logistique	Benktander type II	

TABLE 1.1 : Exemples de distributions appartenant aux trois domaines d'attraction.

La Figure 1.1 présente la densité des trois différents domaines d'attraction.

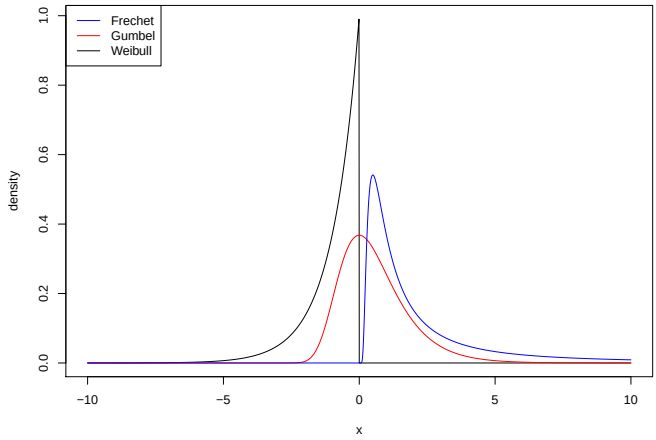

FIGURE 1.1 : Représentation de la densité de chacun des domaines d'attraction, le paramètre de localisation a été fixé à 0.

Parmi les travaux récents sur la théorie des valeurs extrêmes, nous pouvons noter les travaux Worms and Worms [18] et [19], qui s'intéressent à l'estimation de l'indice des valeurs extrêmes d'une distribution à queue lourde dans le cas de présence de censure aléatoire et où la consistence de plusieurs estimateurs est étudiée. Einmahl et al. [20], où les auteurs proposent un estimateur des quantiles extrêmes en présence de censure à droite. Stupfler [21] introduit un estimateur de l'indice des valeurs extrêmes dans les trois domaines d'attraction, la consistence et la normalité asymptotique de cet estimateur sont démontrées. Les auteurs de l'article El Methni et al. [22] établissent les propriétés asymptotiques de "Regression conditional tail moment" (RCTM) ainsi que la normalité asymptotique de l'estimateur à noyau du RCTM. Il est également proposé dans Daouia et al. [23] l'estimation des L^p-quantiles extrêmes et la normalité asymptotique est démontré dans un cadre dépendant.

1.2 Analyse de survie

Dans cette section, nous donnerons les notions essentielles sur la censure. Les estimateurs de Nelson-Aalen (Nelson [7] et Aalen [8]) et de Kaplan-Meier [9] pour les fonctions de survie et de hasard seront ensuite rappeler et nous finirons par détailler le modèle de Cox [6]. Pour plus de détails sur le modèle décrit dans ce chapitre, nous référons à la lecture de Kalbfleisch et Prentice [24], Klein et Moeschberger [25] et Therneau et Grambsch [26].

Soit X une variable aléatoire i.i.d. correspondant au temps de survie de distribu-

tion F et C une variable aléatoire i.i.d. associée au temps de censure de distribution F_C. La notion de la censure est définie dans la Section 1.2.1. On suppose que X et C sont des variables indépendantes entre elles et qu'elles ont chacune une densité positive sur $[x_0, \infty)$, avec $x_0 \geq 0$. On observe l'instant de survie T ainsi que l'indicateur de censure Δ qui sont définis par :

$$T = \min\{X, C\} \qquad \text{and} \qquad \Delta = \mathbb{1}_{X \leq C},$$

où $\mathbb{1}$ est la fonction indicatrice. Nous pouvons remarquer que nous sommes en présence d'un cas de censure à droite. Cela signifie que nous observons le dernier cas enregistré (que ce soit le décès du malade ou la fin de l'expérimentation). Nous notons respectivement par $f(x)$ et $S(x) = 1 - F(x)$ les fonctions de densité et de survie de X. La fonction de hasard $h(x)$ est liée aux fonctions de densité et de survie par les expressions

$$h(x) = f(x)/S(x),$$

et

$$S(x) = e^{-\int_{x_0}^{x} h(u)du}.$$

Un des objectifs de l'analyse de survie est d'estimer la fonction de hasard ou la fonction de survie.

Un travail théorique a été mené dans la thèse de Ndao [4] sur l'estimation des indices de valeurs extrêmes en présence de données censurées. Il a été proposé des estimateurs de l'indice de queue conditionnel et des quantiles extrêmes conditionnels générés à partir de l'estimateur de Weissman adapté et de l'estimateur de Kaplan-Meier. Les propriétés asymptotiques de ces estimateurs ont été déterminées.

1.2.1 La notion de censure

En analyse de survie, il est fréquent d'avoir des observations incomplètes. Il arrive que les données soient collectées partiellement, par exemple cela peut se produire si un patient qui quitte une étude avant la fin de celle-ci.

Definition 1.2.1. *La variable de censure C est définie par la non-observation de l'événement étudié. Si au lieu d'observer X, nous observons C et que nous savons que $Y > C$ ou que $Y < C$ ou encore que $C_1 < Y < C_2$, on dit qu'il y a respectivement censure à droite, censure à gauche ou censure par intervalle.*

Nous pouvons distinguer quatre catégories de censure :

- **Censure à droite :** La variable est dite censurée à droite si nous n'avons aucune information sur sa dernière observation. Nous pouvons donner l'exemple d'un patient malade hospitalisé et nous observons la durée jusqu'au décès ou la guérison de celui-ci. Le cas de censure à droite apparaît si le patient quitte l'hôpital encore malade et que nous n'avons plus de suivi de celui-ci.

- **Censure à gauche :** La variable est dite censurée à gauche lorsqu'un individu a déjà subi l'événement avant d'entrer dans l'étude. Par exemple, si nous souhaitons observer la durée de vie d'un composant électrique. Si nous n'avons pas l'information de la date de mise en place du composant, la durée de vie de ce composant sera censurée à gauche.
- **Censure par intervalle :** Dans ce cas, nous observons une borne inférieure et supérieure à la variable d'intérêt. Par exemple si nous nous intéressons aux patients d'un dentiste possédant un plombage. La variable d'intérêt est le moment où le plombage tombe. La fracture d'un plombage peut être observée uniquement entre deux visites et est donc censurée par intervalle. Certains patients garderont leur plombage à vie (censure à droite) et d'autre où nous connaissons la date exacte car le patient subi une douleur et doit aller chez le dentiste.
- **Censure double ou mixte :** Il y a censure double ou mixte s'il y a présence à la fois de censures à gauche et de censures à droite. Plusieurs modèles non-paramétriques ont été proposés pour l'étude de la censure mixte. On peut notamment citer Turnbull [27] qui est le plus utilisé.

En plus de ces quatre cas de censure, nous pouvons retrouver trois types de censure dans la littérature :

1. **Censure de type I :** L'étude se finie à une date fixée au préalable et nous savons le nombre d'individus dont l'événement n'a pas été observé avant la fin de l'étude. Le nombre d'événements observés est une variable aléatoire. Soit C une valeur fixée correspondant à la fin de l'étude. Si ce sont les variables $Y_1, ..., Y_n$ qui nous intéressent alors nous observons Y_i lorsque $Y_i \leq C$ et C lorsque $Y_i > C$. Ainsi la variable observée est
$$T_i = \min\{Y_i, C\}, \quad \text{pour} \quad i = 1, ..., n.$$
2. **Censure de type II :** L'étude se poursuit jusqu'à ce qu'un nombre d'événements fixé au préalable soit observé. Le nombre d'événement est fixé mais la durée est une variable aléatoire. Par exemple, si nous avons 100 ampoules électriques, nous pouvons arrêter l'étude au moment où 50 ampoules se sont éteintes. Soit $Y_1, ..., Y_n$ la variable d'intérêt ordonnées de sorte que $Y_1 \leq Y_2 \leq ... \leq Y_k \leq ... \leq Y_n$. Si nous gardons k événements, alors nous avons la variable observées :
$$T_1 = Y_1, ..., T_{k-1} = Y_{k-1}, T_k = Y_k, T_{k+1} = Y_k, ..., T_n = Y_k.$$
3. **Censure de type III :** Il s'agit de la censure de type I sauf que la censure est aléatoire. Nous avons donc
$$T_i = \min\{Y_i, C_i\}, \quad \text{pour} \quad i = 1, ..., n.$$

1.2.2 Les estimateurs de Nelson-Aalen et de Kaplan-Meier

Estimer la fonction de hasard

La fonction de hasard joue un rôle majeur dans le modèle de Cox. Il est souvent préférable d'estimer la fonction de hasard cumulée ou intégrée

$$H(x) = \int_{x_0}^{x} h(u)du,$$

que la fonction de hasard. L'estimateur le plus commun de la fonction de hasard cumulée est celui de Nelson-Aalen. Considérons un intervalle de temps λ, nous avons :

$$H(s+\lambda) - H(s) = \mathbb{P}(\text{événements dans } (s, s+\lambda] \,|\, \text{individus à risque au temps } s).$$

Un estimateur naïf de cette probabilité serait $\sum_{i=1}^{n} \mathbb{1}_{s<t_i\leq s+\lambda} / \sum_{i=1}^{n} \mathbb{1}_{s<t_i}$, cela se résume à diviser le nombre d'événements dans la fenêtre $(s, s+\lambda]$ par le nombre d'individus à risque au temps s. Si nous sommons maintenant ces quantités sur l'intervalle $(x_0, t]$, nous obtenons l'estimateur de Nelson-Aalen

$$\widehat{H}(t) = \sum_{\{i\,:t_i\leq t\}} \frac{\sum_{k=1}^{n} \delta_k \mathbb{1}_{t_k=t_i}}{\sum_{j=1}^{n} \mathbb{1}_{t_i<t_j}}. \tag{1.2.1}$$

à partir de l'estimateur (1.2.1), nous pouvons en déduire un estimateur pour la fonction de hasard $\widehat{h}(t) = \frac{\sum_{k=1}^{n} \delta_k \mathbb{1}_{t_k=t}}{\sum_{j=1}^{n} \mathbb{1}_{t<t_j}}$. L'estimateur de Nelson-Aalen est un estimateur de la méthode des moments. Sa variance est estimée uniformément par

$$\text{var}[\widehat{H}(t)] = \sum_{\{i\,:t_i\leq t\}} \frac{\sum_{k=1}^{n} \delta_k \mathbb{1}_{t_k=t_i}}{(\sum_{j=1}^{n} \mathbb{1}_{t_i<t_j})^2}.$$

Cette propriété est justifiée en modélisant le processus de comptage ci-présent par un processus de Poisson où le nombre d'événements dans un intervalle $(t, t+\lambda]$ est approximativement un processus de Poisson pour un h petit. L'estimateur de Nelson-Aalen est consistent uniformément (Andersen et al. [28], Theorem IV.1.1), pour $t < T_{\max}$ nous avons :

$$\sup_{s\in[0,t]} |\widehat{H}(s) - H(s)| \xrightarrow{P} 0, \quad \text{quand} \quad n \to \infty.$$

Estimation la fonction de survie

Pour chaque fonction continue, la fonction de survie et de hasard cumulée sont liées par la relation suivante :

$$S(t) = e^{-H(t)}.$$

Un estimateur de cette fonction de survie suggéré par Breslow [29] est d'inclure l'estimateur de Nelson-Aalen dans la relation précédente, i.e. $\widehat{S}(t) = \widehat{H}(t)$. L'estimateur de Nelson-Aalen se base sur les sauts dans la fonction de hasard à l'instant t, ces sauts s'écrivent de la forme de $\widehat{h}(t)$ et correspondent au nombre d'individus observés à l'instant t divisé par le nombre d'individus à risque à l'instant t. L'estimateur de Breslow peut s'écrire sous la forme :

$$\widehat{S}_B(t) = \prod_{\{j\ :t_j \leq t\}} e^{-\widehat{h}(t)}.$$

L'estimateur de Kaplan-Meier [9] est défini par :

$$\widehat{S}_{KM}(t) = \prod_{\{j\ :t_j \leq t\}} (1 - \widehat{h}(t)).$$

Lorsque les sauts sont petits, nous avons la relation $e^{-\widehat{h}(t)} \simeq 1 - \widehat{h}(t)$. Cette relation tient quand il y a beaucoup de sujets à risque. Une étude a été menée dans le livre Fleming et Harrington [30] pour comparer ces estimateurs pour des échantillons de différentes tailles. En utilisant une étude théorique pour des données non censurées et des simulations pour les données censurées, ils ont montré que l'estimateur de Breslow a une erreur moyenne (MSE) plus petite que l'estimateur de Kaplan-Meier pour t tel que $S(t) \leq 0.2$, mais une MSE plus grande pour $S(t) > 0.2$. Les intervalles de confiance de la fonction de survie peuvent être calculés à partir de l'expression de la variance donné précédemment $\text{var}[\widehat{H}(t)]$, mais aussi de la formule de Greenwood pour la variance du hasard cumulé définie par :

$$\text{var}_g[\widehat{H}(t)] = \int_0^t \frac{\sum_{k=1}^n \delta_k \mathbb{1}_{t_k=t_i}}{(\sum_{j=1}^n \mathbb{1}_{t_i<t_j})(\sum_{j=1}^n \mathbb{1}_{t_i<t_j} - \sum_{k=1}^n \delta_k \mathbb{1}_{t_k=t_i})}.$$

Les intervalles de confiance sont calculés pour $S(t)$ par $\widehat{S}_{KM}(t) \pm 1.96\sqrt{\text{var}[\widehat{S}_{KM}(t)]}$ où $\text{var}[\widehat{S}_{KM}(t)] \simeq \widehat{S}^2_{KM}(t)\text{var}[\widehat{H}(t)]$. Dans l'expression précédente, nous pouvons remplacer $\text{var}[\widehat{H}(t)]$ par $\text{var}_g[\widehat{H}(t)]$ pour l'estimateur de Kaplan-Meier.

L'estimateur de Kaplan-Meier est asymptotiquement gaussien, nous avons le résultat suivant :

Theorem 1.2.2. *(Droesbeke et Saporta [31] 2011)*

Si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors quand $n \to \infty$,

$$\sup_{t\geq 1} |\widehat{S}_{KM}(t) - S(t)| \overset{p.s.}{\to} 0.$$

Pour tout $t \geq 0$, *on a quand* $n \to \infty$:

$$\sqrt{n}(\widehat{S}_{KM}(t) - S(t)) \overset{d}{\to} W_t,$$

où $(W_t)_{t\geq 0}$ *est un processus gaussien centré qui vérifie pour tout* t *et* s *strictement positifs*

$$Cov(W_s, W_t) = S(t)S(s)\int_0^{\min\{t,s\}} \frac{f(u)}{S(u)^2(1-F_C(u))}du$$

L'estimateur de Nelson-Aalen est défini au-delà de la dernière observation de la façon suivante : si la dernière observation n'est pas censurée, alors la valeur de l'estimateur est nulle. Si la dernière observation est censurée, alors la valeur de l'estimateur devient constante et positive. L'estimateur est donc légèrement biaisé sur la queue de la distribution, $\widehat{H}(t)$ reste constant quand t est supérieur à la dernière observation alors que nous pouvons suggérer que $H(t)$ continue d'augmenter. Nous pouvons également estimer la fonction de hasard cumulée par $-\ln(\widehat{S}_{KM}(t))$. Cet estimateur a une estimation consistante de sa variance par la formule de Greenwood $\text{var}_g[\widehat{H}(t)]$. Lorsque la dernière observation est non censurée, le logarithme de l'estimateur de Kaplan-Meier est infini et n'est pas un bon estimateur pour la fonction de hasard cumulée dans la queue de la distribution.

1.2.3 Modèle de Cox

Le modèle de Cox [6] est devenu la procédure la plus utilisée pour modéliser les relations entre les covariables et la fonction de survie avec des données censurées. Soit Z une variable aléatoire correspondant aux covariables. On suppose que X et C sont indépendantes conditionnellement à Z. Lorsque les covariables sont considérées fixes dans le temps, le modèle est similaire à une régression linéaire multiple.

Le modèle de Cox introduit la relation suivante :

$$h(\,t\,|\,z\,) = h_0(t)e^{\beta\cdot z},$$

où β est un vecteur de paramètres et h_0 est une fonction positive inconnue appelée "baseline hazard". Le modèle est souvent appelé modèle à hasards proportionnels.

L'estimation de β est basée sur la fonction de vraisemblance partielle introduite par Cox [6]. Cette fonction à la forme

$$PL(\beta) = \prod_{i=1}^{n}\prod_{t\geq 0}\left(\frac{\mathbb{1}_{t_i\geq t}e^{\beta\cdot z_i}}{\sum_{j=1}^{n}\mathbb{1}_{t_j\geq t}e^{\beta\cdot z_j}}\right)^{\delta_i}.$$

La fonction de log-vraisemblance partiel peut s'écrire comme la somme :

$$LV(\beta) = \sum_{i=1}^{n}\int_0^{\infty}\left[\mathbb{1}_{t_i\geq t}\beta\cdot z_i - \ln\left(\sum_{j=1}^{n}\mathbb{1}_{t_j\geq t}e^{\beta\cdot z_j}\right)\right]\delta_i dt.$$

Dérivant la fonction de log-vraisemblance partiel en fonction de β donne le vecteur de score $U(\beta)$ défini par :

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left[z_i - \frac{\sum_{j=1}^n \mathbb{1}_{t_j \geq t} e^{\beta \cdot z_j} z_j}{\sum_{j=1}^n \mathbb{1}_{t_j \geq t} e^{\beta \cdot z_j}} \right] \delta_i dt.$$

L'estimateur du maximum de vraisemblance se trouve en résolvant l'équation

$$U(\widehat{\beta}) = 0.$$

Nous avons le résultat suivant pour l'estimateur de β (Therneau et Grambsch [26])

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathbb{E}(\mathcal{I}(\beta))}\right),$$

où $\mathcal{I}(\beta)$ est la matrice d'information de Fischer trouvée par la dérivée seconde de $LV(\beta)$:

$$\mathcal{I}(\beta) = \sum_{i=1}^{n} \int_0^\infty V(\beta, t) \delta_i dt,$$

et

$$V(\beta, t) = \frac{\sum_{i=1}^n \mathbb{1}_{t_j \geq t} e^{\beta \cdot z_j} \left[z_i - \frac{\sum_{j=1}^n \mathbb{1}_{t_j \geq t} e^{\beta \cdot z_j} z_j}{\sum_{j=1}^n \mathbb{1}_{t_j \geq t} e^{\beta \cdot z_j}} \right]' \left[z_i - \frac{\sum_{j=1}^n \mathbb{1}_{t_j \geq t} e^{\beta \cdot z_j} z_j}{\sum_{j=1}^n \mathbb{1}_{t_j \geq t} e^{\beta \cdot z_j}} \right]}{\sum_{j=1}^n \mathbb{1}_{t_j \geq t} e^{\beta \cdot z_j}}.$$

En pratique, l'algorithme de Newton-Raphson peut est utilisé pour résoudre l'équation $U(\widehat{\beta}) = 0$.

1.3 Modèle linéaire et chaîne de Markov à états cachés

Le modèle présenté dans cette section est l'objet de l'article Durand et al. [32]. Celui-ci s'intéresse à la modélisation et l'analyse de courbes de consommation électrique. Dans un premier temps, une analyse de la variance est effectuée sur la log-consommation électrique puis les erreurs sont modélisées par une chaîne de Markov à états cachés. Dans l'article, ces états sont interprétés et mis en relation avec la consommation de différents appareils lorsque l'information est disponible. Dans cette section, nous présentons un modèle plus simple que celui utilisé dans Durand et al. [32] car le but est de modéliser individuellement la courbe de consommation électrique.

1.3.1 Modèle linéaire

Pour pouvoir modéliser la courbe de consommation par un modèle linéaire, nous considérons que la consommation électrique d'un foyer s'explique par des périodes

homogènes et que les fluctuations à l'intérieur de ces périodes sont dues au hasard. Nous considérons le modèle linéaire pour le logarithme de la consommation électrique individuelle.

$$\ln(G_t) = \alpha + \beta C_t + \gamma T_t + \epsilon_t, \tag{1.3.1}$$

où G_t est la courbe de charge (ou courbe de consommation) décrite à la Section 1.4.1 au temps $t \in [t_{min}\,;t_{max}]$, α est une constante, γ correspond au lien linéaire entre la température extérieure (T_t) et la courbe de charge, β est un vecteur de paramètre représentant les différents liens avec la variable chronologique, C_t est la variable chronologique et est composée de l'information sur :

- l'heure
- le jour de la semaine (week-end ou semaine de travail)
- le mois
- l'année
- interaction entre le mois et l'année
- interaction entre le mois et l'heure.

Enfin, ϵ_t caractérisent les erreurs du modèle qui sont une suite de variables aléatoires. En pratique, les erreurs (ϵ_t) sont souvent une suite de variables aléatoires indépendantes et gaussiens de moyenne nulle. Cette hypothèse n'est pas vérifiée avec les données présentées Section 1.4.1. Il est proposé dans Durand et al. [32] de modéliser le processus des (ϵ_t) par une chaîne de Markov à états cachés.

1.3.2 Chaîne de Markov à états cachés

L'idée derrière la modélisation des erreurs par une chaîne de Markov à états cachés est que la consommation électrique d'un logement peut être considérée comme des périodes homogènes pendant lesquelles la consommation fluctue en fonction de comportements. Ceux-ci peuvent être interprétés comme une activité domestique (télévision, lave-linge, etc...).

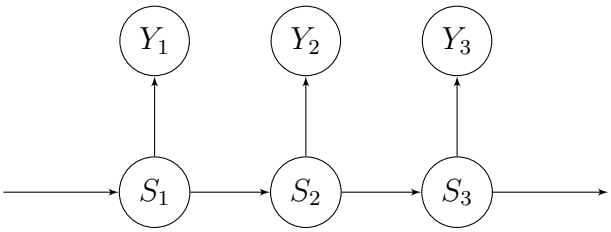

FIGURE 1.2 : Graphe d'une chaîne de Markov à états cachés basique.

La Figure 1.2 représente la structure d'un modèle simple de chaîne de Markov à états cachés. Soit $(Y_1, ..., Y_n)$ un processus correspondant à la variable observée et $(S_1, ..., S_n)$ un processus caché correspondant aux états cachés à valeurs dans $\{1, ..., K\}$, où K est le nombre d'états cachés du modèle. Nous considérons que

- $S_1, ..., S_n$ est une chaîne de Markov homogène, irréductible et stationnaire de matrice de transition P et de distribution stationnaire $\pi = (\pi_1, ..., \pi_K)$ avec $\pi_i \geq 0$, pour $i = 1, ..., K$ et $\sum_{i=1}^{K} \pi_i = 1$.
- Les Y_t sont indépendants conditionnellement aux S_t.
- Sachant $S_t = j$, Y_t suit une loi de densité f_{θ_j} appartenant à une famille paramétrée $(f_\theta)_{\theta \in \Theta}$.

Il s'ensuit que la loi de $(Y_1, ..., Y_n)$ suit une chaîne de Markov à états cachés avec les paramètres $\lambda = (\pi, P, \theta_1, ..., \theta_K)$. Le choix de la famille de distribution $(f_\theta)_{\theta \in \Theta}$ est important car il permet de modéliser la nature du bruit. Le choix d'une famille gaussienne, de telle sorte que $\theta = (\mu, \sigma^2)$ où μ et σ^2 dénotent respectivement la moyenne et la variance, est motivé par le fait que les erreurs semblent suivre un mélange gaussien comme le montre l'histogramme de la densité Figure 1.3. Ce fait a été validé dans Durand et al. [32] et par les données que nous avons.

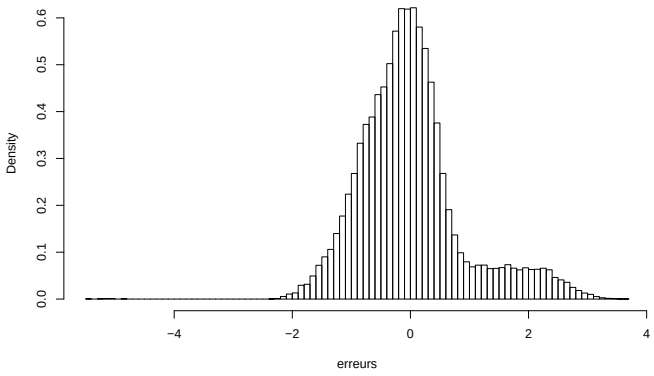

FIGURE 1.3 : Histogramme de la densité des erreurs du modèle 1.3.1 pour un foyer pris au hasard.

Les paramètres θ sont ensuite estimés par la méthode du maximum de vraisemblance en utilisant l'algorithme EM.

L'algorithme EM ("Expectation-Maximization") est une méthode permettant de trouver l'estimateur du maximum de vraisemblance d'un ou de plusieurs paramètres d'une distribution sachant les données observées. Pour plus d'information concernant

cette méthode, nous référons à Dempster et al. [33], Redner et Walker [34], Ghahramani et Jordan [35], Jordan et Jacobs [36], Bishop et al. [37] et Wu [38]. L'algorithme EM est utilisé pour deux sortes d'applications, la première est lorsque nous somme en présence de données manquantes. La deuxième est lors de l'optimisation de la fonction de vraisemblance lorsque celle-ci ne peut être résolue analytiquement, ce qui est le cas ici. Ci-dessous, nous présentons rapidement l'algorithme EM, pour plus de détails sur la mise en place de cet algorithme pour l'estimation des paramètres d'un mélange gaussien, nous référons à Bilmes et al. [39].

Soit $\mathcal{X}$ les données observées et $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ le set complet des données avec la fonction de densité jointe :

$$p(z \mid \Theta) = p(x, y \mid \Theta) = p(y \mid x, \Theta) p(x \mid \Theta).$$

Grâce à cette fonction, nous pouvons définir une nouvelle fonction de vraisemblance, $\mathcal{L}(\Theta \mid \mathcal{Z}) = p(\mathcal{X}, \mathcal{Y} \mid \Theta)$. La première étape de l'algorithme EM, soit "Expectation", consiste à trouver l'espérance de la fonction $\ln p(\mathcal{X}, \mathcal{Y} \mid \Theta)$ en fonction de $\mathcal{Y}$ sachant les données observées $\mathcal{X}$ et l'estimation actuelle des paramètres.

$$\mathcal{Q}(\Theta, \Theta^{(i-1)}) = \mathbb{E}\left[\ln p(\mathcal{X}, \mathcal{Y} \mid \Theta) \,\middle|\, \mathcal{X}, \Theta^{(i-1)} \right],$$

où $\Theta^{(i-1)}$ sont l'estimation actuelle des paramètres que nous usons pour évaluer l'espérance et Θ sont les nouveaux paramètres que nous optimisons pour augmenter $\mathcal{Q}$.

La seconde étape de l'algorithme EM, soit "Maximization", consiste à maximiser l'espérance que nous avons calculé lors de la première étape :

$$\Theta^{(i)} = \underset{\Theta}{argmax} \mathcal{Q}(\Theta, \Theta^{(i-1)}).$$

La répétition des deux étapes est nécessaire et permet d'augmenter la valeur de la log-vraisemblance pour converger vers un maximum local.

Un autre paramètre à prendre en compte est celui du nombre d'états K. Une étude a été menée dans Durand et al. [32] pour définir combien d'états nous allons choisir pour la chaîne de Markov à états cachés. Ce nombre d'état a été choisi par les critères de sélection BIC, ICL (Biernacki et al. [40]) et le demi-échantillonnage. De plus, le tableau d'intensité en fonction du nombre d'état a été effectué en calculant l'indice descriptif du Φ^2 normalisé. Cet indice mesure l'écart des données à l'indépendance entre les douze usages et des K états. En définitif, le modèle le plus simple avec 7 états a été sélectionné parmi les modèles étudiés.

1.4 Description des données

Dans cette partie, nous décrirons les données sur lesquelles les résultats ont été produits. Nous allons commencer par les données de consommation électrique collectées par le compteur Linky. Les données d'informations accompagnant celles de la consommation électrique sont également expliquées. Nous présenterons enfin les données recueillies par les enquêtes ainsi que par les stations météo.

1.4.1 Données de consommation électrique

Ces données se composent de courbes de charge et d'index. Les données de courbes de charge représentent l'appel de puissance moyen entre deux mesures (en Watt) et une mesure est effectuée toutes les 10 minutes. Les données d'index sont les consommations journalières cumulées d'un foyer, nous disposons d'une mesure tous les jours (en Watt-heure). Les données d'index et de courbe de charge sont liées entre elles, en effet les index correspondent à l'intégrale cumulée de la courbe de charge (ou la somme dans notre cas). Nous appellerons par la suite les données de courbes de charge la puissance appelée moyenne PAM. Les données d'index seront spécifiquement citées.

Pour des raison techniques, le déploiement des compteurs Linky a démarré en juillet 2015 et cela a retardé le recueil des données qui a commencé fin décembre 2015.

Courbes de charge

La Figure 1.4 permet d'avoir un aperçu d'une courbe de charge sur une journée pour un foyer. On peut noter que la consommation électrique fluctue en fonction des divers usages du foyers que nous ne connaissons pas.

Nous allons regarder l'évolution du nombre de foyers sur la période de l'étude. La Figure 1.5 permet de regarder le nombre de foyers avec respectivement au moins 10%, 30%, 50%, 70% et 90% des données disponibles. Par exemple sur le mois de janvier 2016, si nous disposions de toutes les mesures pour un foyer, nous aurions 4464 mesures (6 (mesures dans une heure) $\times$24 (heures) $\times$31 (jours)). Sur ce même mois, nous avons 354 foyers avec au moins $4464 \times 0.1 = 446.4$ mesures dans le mois (rouge sur la figure) et 251 foyers avec au moins $4464 \times 0.9 = 4017.6$ mesures (vert sur la figure).

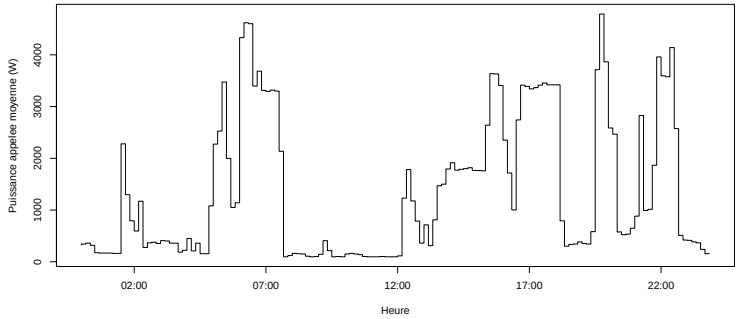

FIGURE 1.4 : Courbe de charge d'un foyer avec une puissance souscrite de 9 kVA pour une journée.

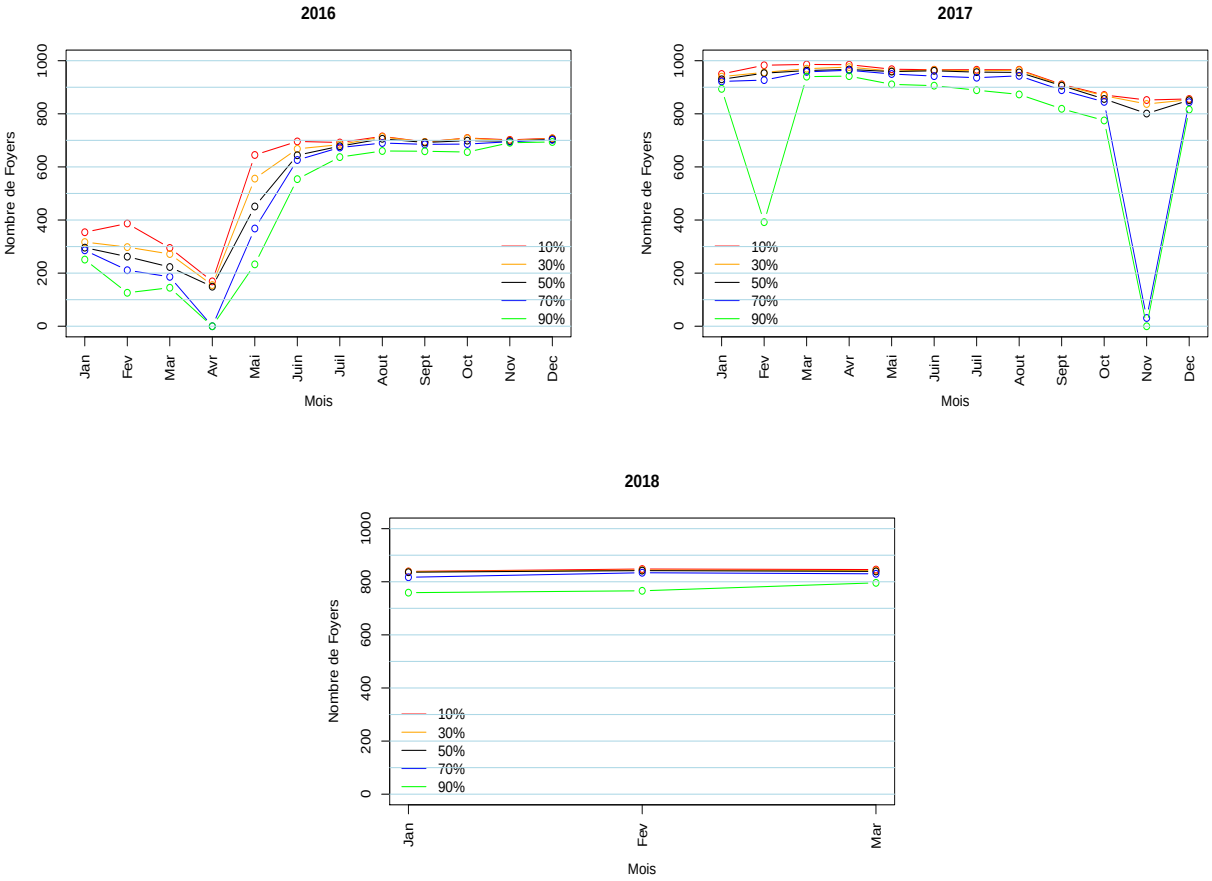

FIGURE 1.5 : Évolution du nombre de foyers dont nous possédons les courbes de charge.

Nous pouvons observer que le nombre de foyers dont nous possédons les mesures reste constant et élevé à partir de juillet 2016. Nous pouvons tout de même remarquer

qu'une baisse de la qualité a été observée en novembre 2017. Cette baisse est due à une baisse de vigilance qui a été réglée rapidement.

Index

Les données d'index sont le cumul de la consommation journalière. Pour avoir la consommation journalière non cumulée, il suffit de différencier les données. La Figure 1.6 trace la courbe d'index d'un foyer sur plusieurs jours (non cumulée).

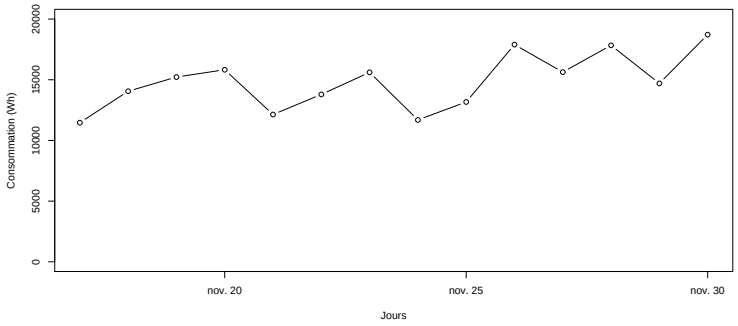

FIGURE 1.6 : Exemple de la consommation électrique d'un foyer avec une puissance souscrite de 9 kVA sur plusieurs jours.

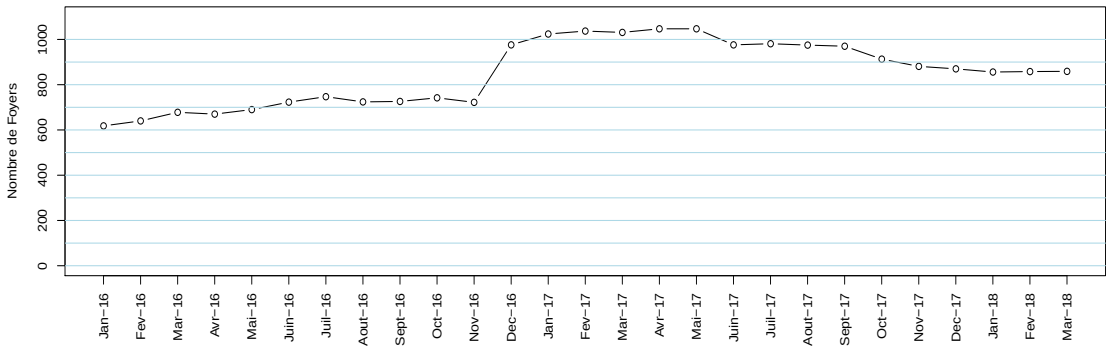

FIGURE 1.7 : Évolution du nombre de foyers dont nous possédons les index.

La Figure 1.7 nous permet de regarder l'évolution du nombre de foyers sur la durée du projet pour chaque mois avec au moins une mesure d'index.

On peut remarquer que le nombre de foyers avec des mesures d'index reste constant sur la période du projet. L'augmentation en décembre 2016 - janvier 2017 est due à un recrutement supplémentaire de foyers.

Panel d'appartenance et puissance souscrite

Nous disposons d'un fichier accompagnant les données de consommation et renseignant le type de contrat souscrit pour chaque foyer. Nous avons 22 foyers qui ont changé de contrat en cours de projet, nous avons donc retiré ces foyers de l'étude.

Nous avons deux informations importantes sur le type de contrat :

- Puissance souscrite : La puissance maximale autorisée pour le foyer. Si la puissance appelée dépasse celle-ci, alors le foyer subit une ouverture de breaker.
- HP/HC : Il s'agit d'un contrat heure pleine/heure creuse où le tarif de l'électricité varie en fonction de l'heure.

En 2018, la Table 1.2 montre la répartition des foyers selon leur type de contrat :

Puissance souscrite	Nombre de foyers contrat base	Nombre de foyers contrat HP/HC	Total
3 kVA	33	0	33
6 kVA	384	138	522
9 kVA	30	197	227
12 kVA	1	45	46
15 kVA	0	1	1
18 kVA	2	5	7
Total	450	386	836 foyers

TABLE 1.2 : Répartition des foyers en fonction du contrat de puissance souscrite.

Nous regardons maintenant l'appartenance des foyers à un panel. Lors de l'inscription d'un foyer au projet SOLENN, celui-ci a été assigné à l'un des panels suivants :

- Individuel ALOEN : foyers suivis individuellement par ALOEN et ayant eu des visites.
- Collectif ALOEN (hiv. 1) : foyers suivis collectivement par ALOEN et des ateliers collectifs ont été organisés. Ces foyers ont été recrutés en 2015.
- Collectif ALOEN (hiv. 2) : foyers ayant été recrutés lors de la deuxième vague de recrutement en 2016/2017.
- Vity : foyers ayant été suivis par l'entreprise Vity et un appareil domotique a été installé chez eux.

- Delta Dore : foyers ayant été suivis par l'entreprise Delta Dore et un appareil domotique a été installé chez eux.

- Témoins : foyers n'ayant participé à aucune activité.

La modulation de puissance (MP) correspond à l'écrêtement.

Panel / MP	Individuel ALOEN	Collectif (hiv. 1) ALOEN	Collectif(hiv. 2) ALOEN	Vity
Sans MP	27	33	164	23
Avec MP	26	43	97	23
Total	53	76	261	46
	Delta Dore	Témoin	Sortie Expé. ou autre	Total
Sans MP	11	238		496
Avec MP	9	86		284
Total	20	324	56	836

TABLE 1.3 : Nombre de foyers dans chaque panel.

En observant les deux tableaux précédents, on remarque que le nombre de foyers est faible si nous séparons par panels ainsi que par puissance souscrite.

Écrêtement

Plusieurs informations concernant les tirs d'écrêtement ont été recueillies. La date et durée des écrêtements pendant l'hiver 2017/2018 ainsi que le pourcentage d'écrêtement sont présentés à la Table 1.4. Il y a eu 22 tirs d'écrêtements répartis sur les trois hivers. Le talon représente la puissance maximale où l'écrêtement est possible, cela veut dire que si la puissance en période d'écrêtement théorique est inférieure au talon, alors la puissance maximale devient le talon.

Période d'écrêtement	Diminution (talon)	3 kVA	6 kVA	9 kVA	12 kVA
9\11\2017 (18-20h)	70% (1.8 kVA)	1.8	1.8	2.7	3.6
29\11\2017(18-20h)	80% (1.8 kVA)	1.8	1.8	1.8	2.4
19\12\2017 (17-19h)	80% (2 kVA)	2	2	2	2.4
8\1\2018 (18-20h)	80% (2 kVA)	2	2	2	2.4
2\2\2018 (11-13h)	80% (2 kVA)	2	2	2	2.4
20\2\2018 (18-20h)	80% (2 kVA)	2	2	2	2.4
13\3\2018 (18-20h)	70% (2 kVA)	2	2	2.7	3.6
28\3\2018 (18-20h)	81% (2 kVA)	2	2	2	2.376

TABLE 1.4 : Récapitulatif des écrêtements et de la puissance maximale (en kVA) durant les périodes d'écrêtement.

De plus, les information sur toutes les ouvertures de breaker (quand un foyer disjoncte et n'a plus d'électricité) ainsi que sur la période pendant laquelle le foyer se retrouve sans électricité sont également fournies.

1.4.2 Enquêtes

Nous disposons des données de trois enquêtes effectuées durant la période du projet. Ces enquêtes ont eu lieu en novembre 2015, juin 2016 et mai - juin 2017. Il a été en mesure de collecter 375 questionnaires sur l'ensemble des foyers interrogés. Plusieurs données d'informations sur le logement ont été collectées sur ces foyers tel que l'information sur le type de logement, le mode de chauffage principal, la surface des logements, l'âge des répondants, le nombre de personnes dans le foyers, etc... .

1.4.3 Météo

Des données météo ont été recueillies pendant la durée du projet. Des stations météo ont été posées chez certains foyers volontaires et les données de température intérieure et extérieure ainsi que diverses mesures ont été recueillies sur les serveurs de l'UBS, le pas de temps entre chaque mesure est de 5 minutes. Nous nous sommes servis uniquement des données de températures extérieures car elles sont communes à tous les foyers, même pour ceux n'ayant pas de station météo. À partir des données de température extérieures, nous avons construit une température commune pour chacun des foyers en prenant la médiane des données mesurées à chaque instant. Cette information est importante car nous avons un impact de la température extérieure sur la consommation électrique du foyer, surtout quand celui-ci dispose d'un chauffage électrique. La Figure 1.8 montre la médiane des températures exté-

rieures collectées sur tous les foyers possédant une station météo pour une semaine de janvier 2017.

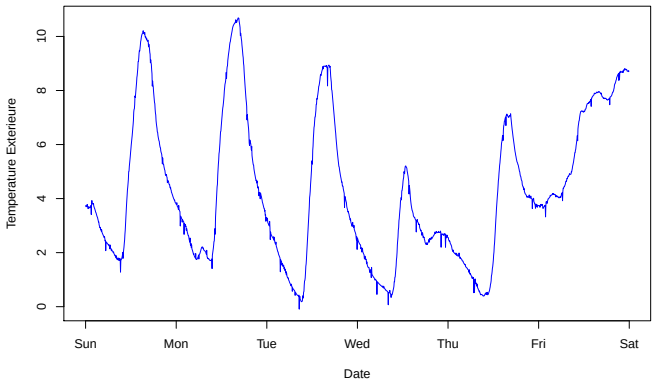

FIGURE 1.8 : Température extérieure pour une semaine de janvier 2017.

Chapter 2

extremefit: An R package for extreme quantiles

This chapter is the subject of the article Durrieu et al. [41] in collaboration with Gilles Durrieu, Ion Grama, Quang-Khoai Pham and Jean-Marie Tricot to appear in *Journal of Statistical Software*.

RÉSUMÉ. Le package **extremefit** permet d'estimer des quantiles extrêmes et des probabilités d'événements rares. L'idée de l'approche est d'ajuster la queue de la fonction de distribution au-delà d'un seuil par une loi de Pareto. Nous proposons une procédure dépendante des données pour choisir le seuil. Un exemple sur données simulées ainsi que deux exemples sur des données réelles sont donnés.

ABSTRACT. **extremefit** is a package to estimate the extreme quantiles and probabilities of rare events. The idea of our approach is to adjust the tail of the distribution function over a threshold with a Pareto distribution. We propose a pointwise data driven procedure to choose the threshold. To illustrate the method, we use simulated data sets example and two real-world data sets available on the package.

2.1 Introduction

Extreme values study plays an important role in several practical domains of applications, such as insurance, biology and geology. For example, in Buishand et al. [10], the authors study extremes to determine how severe the rainfall periods occur in North Holland. Sharma et al. [42] use extreme values procedure to predict violations of air quality standards. Various applications were presented in a lot of areas such as hydrology (Davison and Smith [43] and Katz et al. [44]), insurance ([16] and Rootzén and Tajvidi [17]) or finance (Danielsson and de Vries [12], McNeil [13], Embrechts et al. [45] and Gencay and Selcuk [15]). Other applications goes from rainfall data (Gardes and Girard [11]) to earthquake (Sornette et al. [46]). The extreme value theory consists using appropriate statistical models to estimate extreme quantiles and probability of rare events.

The idea of the approach implemented in the **extremefit** package is to fit a Pareto distribution to the data over a threshold τ using the Peak-Over-Threshold method. The choice of τ is a challenging problem, a large value can lead to an important variability while a small value may increase the bias. We refer to Hall and Welsh [47], Drees and Kaufmann [48], Guillou and Hall [49], Huisman et al. [50], Beirlant et al. [5], Grama and Spokoiny [51, 52] and El Methni et al. [53] where several procedures for choosing the threshold τ have been proposed. Here, we adopt the method from Grama and Spokoiny [51] and Durrieu et al. [54]. The package **extremefit** includes the modeling of time dependent data. The analysis of time series involves a bandwidth parameter h whose data driven choice is non trivial. We refer to Staniswalis [55] and Loader [56] for the choice of the bandwidth in a nonparametric regression. For the purposes of extreme value modeling, we use a cross-validation approach from Durrieu et al. [54].

The **extremefit** package for the R system ([57]) is based on the methodology described in Durrieu et al. [54]. The package performs a nonparametric estimation of extreme quantiles and probabilities of rare events. It proposes a pointwise choice of the threshold τ and, for the time series, a global choice of the bandwidth h and gives graphical representations of the results.

The chapter is organized as follows. Section 2.2 is an overview of several existing R packages dealing with extreme value analysis. In Section 2.3, we describe the model and the estimation of the parameters, including the threshold τ and the bandwidth h choices. Section 2.4 contains a simulation study whose aim is to illustrate the performance of our approach. In Section 3.8, we give several applications on real data sets and we conclude in Section 2.6.

2.2 Extreme value packages

There exist several R packages dealing with the extreme value analysis. We give a short description of some of them. For a detailed description of these packages, we refer to Gilleland et al. [58]. A CRAN task view exists in extreme value analysis giving a description of registered packages available in CRAN, see https://CRAN.R-project.org/view=ExtremeValue. Among those available packages, the well known Peak-Over-Threshold method we mentioned before, has many developments especially illustrated by the **POT** package, https://CRAN.R-project.org/package=POT (Ribatet and Dutang [59]).

Some of the packages have a specific use, such as the package **SpatialExtremes** (Ribatet et al. [60]), which models spatial extremes and provides maximum likelihood estimation, bayesian hierarchical and copula modeling, or the package **fExtremes** (Wuertz and many others [61]) for financial purposes using functions from the packages **evd**, **evir** and others.

The **copula** package (Hofert et al. [62]) provides tools for exploring and modeling dependent data using copulas. The **evd** package (Stephenson and Ferro [63]) provides both block maxima and peak-over-threshold computations based on maximum likelihood estimation in the univariate and bivariate cases. The **evdbayes** package (Stephenson and Ribatet [64]) provides an extension of the **evd** package using bayesian statistical methods for univariate extreme value models. The package **extRemes** (Gilleland [65]) implements also univariate estimation of block maxima and peak-over-threshold by maximum likelihood estimation allowing non stationarity. The package **evir** (Pfaff et al. [66]) is based on fitting a generalized Pareto distribution with the Hill estimator over a given threshold. The package **lmom** (Hosking [67]) is dealing with L-moments to estimate the parameters of extreme value distributions and quantile estimations for reliability or survival analysis. The package **texmex** (Southworth and Heffernan [68]) provides statistical extreme value modeling of threshold excesses, maxima and multivariate extremes, including maximum likelihood and Bayesian estimation of parameters.

In contrast to previous described packages, the **extremefit** package provides tools for modeling heavy tail distributions without assuming a general parametric structure. The idea is to fit a parametric Pareto model to the tail of the unknown distribution over some threshold. The remaining part of the distribution is estimated nonparametrically and a data driven algorithm for choosing the threshold is proposed in Section 2.3.2. We also provide a version of this method for analyzing extreme values of a time series based on the nonparametric kernel function estimation approach. A data driven choice of the bandwidth parameter is given in Section 2.3.3. These estimators are studied in more details in Durrieu et al. [54].

2.3 Extreme value prediction

2.3.1 Model and estimator

We consider $F_t(x) = P(X \leq x|T = t)$ the conditional distribution of a random variable X given a time covariate $T = t$, where $x \in [x_0, \infty)$ and $t \in [0, T_{\max}]$. We observe independent random variables $X_{t_1}, ..., X_{t_n}$ associated to a sequence of times $0 \leq t_1 < ... < t_n \leq T_{\max}$, such that for each t_i, the random variable X_{t_i} has the distribution function F_{t_i}. The purpose of the **extremefit** package is to provide a pointwise estimation of the tail probability $S_t(x) = 1 - F_t(x)$ and the extreme p-quantile $F_t^{-1}(p)$ processes for any $t \in [0, T_{\max}]$, given $x > x_0$ and $p \in (0, 1)$. We assume that F_t, are in the domain of attraction of the Fréchet distribution. The idea is to adjust, for some $\tau \geq x_0$, the excess distribution function

$$F_{t,\tau}(x) = 1 - \frac{1 - F_t(x)}{1 - F_t(\tau)}, \quad x \in [\tau, \infty) \tag{2.3.1}$$

by a Pareto distribution:

$$G_{\tau,\theta}(x) = 1 - \left(\frac{x}{\tau}\right)^{-\frac{1}{\theta}}, \quad x \in [\tau, \infty), \tag{2.3.2}$$

where $\theta > 0$ and $\tau \geq x_0$ an unknown threshold, depend on t. The justification of this approach is given by the Fisher-Tippett-Gnedenko theorem (Beirlant et al. [5], Theorem 2.1) which states that F_t, is in the domain of attraction of the Fréchet distribution if and only if $1 - F_{t,\tau}(\tau x) \to x^{-1/\theta}$ as $\tau \to \infty$. This consideration is based on the peak-over-threshold (POT) approach (Beirlant et al. [5]). We consider the semi-parametric model defined by:

$$F_{t,\tau,\theta}(x) = \begin{cases} F_t(x) & \text{if } x \in [x_0, \tau], \\ 1 - (1 - F_t(\tau))(1 - G_{\tau,\theta}(x)) & \text{if } x > \tau, \end{cases} \tag{2.3.3}$$

where $\tau \geq x_0$ is the threshold parameter. We propose in the sequel to estimate F_t and θ which are unknown in (2.3.3).

The estimator of $F_t(x)$ is taken as the weighted empirical distribution given by

$$\widehat{F}_{t,h}(x) = \frac{1}{\sum_{j=1}^{n} W_{t,h}(t_j)} \sum_{i=1}^{n} W_{t,h}(t_i) \mathbb{1}_{\{X_{t_i} \leq x\}}, \tag{2.3.4}$$

where, for $i = 1, \ldots, n$, $W_{t,h}(t_i) = K\left(\frac{t_i - t}{h}\right)$ are the weights and $K(.)$ is a kernel function assumed to be continuous, non-negative, symmetric with support on the real line such that $K(x) \leq 1$, and $h > 0$ is a bandwidth.

By maximizing the weighted quasi-log-likelihood function (see Durrieu et al. [54], Staniswalis [55] and Loader [56])

$$\mathcal{L}_{t,h}(\tau, \theta) = \sum_{i=1}^{n} W_{t,h}(t_i) \ln \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}) \tag{2.3.5}$$

with respect to θ, we obtain the estimator

$$\widehat{\theta}_{t,h,\tau} = \frac{1}{\widehat{n}_{t,h,\tau}} \sum_{i=1}^{n} W_{t,h}(t_i) \mathbb{1}_{\{X_{t_i} > \tau\}} \ln\left(\frac{X_{t_i}}{\tau}\right), \tag{2.3.6}$$

where $\widehat{n}_{t,h,\tau} = \sum_{i=1}^{n} W_{t,h}(t_i) \mathbb{1}_{\{X_{t_i} > \tau\}}$ is the weighted number of observations over the threshold τ.

Plug-in (2.3.4) and (2.3.6) in the semi-parametric model (2.3.3), we obtain:

$$\widehat{F}_{t,h,\tau}(x) = \begin{cases} \widehat{F}_{t,h}(x) & \text{if } x \in [x_0, \tau], \\ 1 - \left(1 - \widehat{F}_{t,h}(\tau)\right)\left(1 - G_{\tau,\widehat{\theta}_{t,h,\tau}}(x)\right) & \text{if } x > \tau. \end{cases} \tag{2.3.7}$$

For any $p \in (0,1)$, the estimator of the p-quantile of X_t is defined by

$$\widehat{q}_p(t,h) = \begin{cases} \widehat{F}_{t,h}^{-1}(p) & \text{if } p < \widehat{p}_\tau, \\ \tau\left(\frac{1-\widehat{p}_\tau}{1-p}\right)^{\widehat{\theta}_{t,h,\tau}} & \text{otherwise,} \end{cases} \tag{2.3.8}$$

where $\widehat{p}_\tau = \widehat{F}_{t,h}(\tau)$.

2.3.2 Selection of the threshold

The determination of the threshold τ in the model (2.3.3) is based on a testing procedure which is a goodness-of-fit test for the parametric-based part of the model. At each step of the procedure, the tail adjustment to a Pareto distribution is tested based on k upper statistics. If it is not rejected, the number k of upper statistics is increased and the tail adjustment is tested again until it is rejected. If the test rejects the parametric tail fit from the very beginning, the Pareto tail adjustment is not significant. On the other hand, if all the tests accept the parametric Pareto fit then the underlying distribution is significantly Pareto. The critical value denoted by D depends on the kernel choice and is determined by Monte-Carlo simulation, using the `CriticalValue` function of the package.

In Table 2.1, we display the critical values using `CriticalValue` obtained for several kernel functions.

The default values of the parameters in the algorithm yield satisfying estimation results on simulation study without being time-consuming even for large data sets. The choice of these tuning parameters is given in Durrieu et al. [54].

The following commands compute the critical value D for the truncated Gaussian kernel with $\sigma = 1$ (default value) and display the empirical distribution function of the goodness-of-fit test statistic which determines the threshold τ. The parameters n and N_{MC} defines respectively the sample size and the number of Monte-Carlo simulated samples.

Kernel	D	$K(x)$
Biweight	7	$\frac{15}{16}(1-x^2)^2\mathbb{1}_{\lvert x\rvert\leq 1}$
Epanechnikov	6.1	$\frac{3}{4}(1-x^2)\mathbb{1}_{\lvert x\rvert\leq 1}$
Rectangular	10.0	$\mathbb{1}_{\lvert x\rvert\leq 1}$
Triangular	6.9	$(1-\lvert x\rvert)\mathbb{1}_{\lvert x\rvert\leq 1}$
Truncated Gaussian, $\sigma = 1/3$	8.3	$\frac{3}{\sqrt{2\pi}}\exp(-\frac{(3x)^2}{2})\mathbb{1}_{\lvert x\rvert\leq 1}$
Truncated Gaussian, $\sigma = 1$	3.4	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathbb{1}_{\lvert x\rvert\leq 1}$

Table 2.1: Critical values associated to kernel functions. The Gaussian kernel with standard deviation 1/3 is approximated by the truncated Gaussian kernel $\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)\mathbb{1}_{|x|\leq 1}$ with $\sigma = 1/3$.

```
R> library("extremefit")
R> n <- 1000
R> NMC <- 500
R> CriticalValue(NMC, n, TruncGauss.kernel, prob = 0.99, plot = TRUE)

[1] 3.432665
```

For a given t, the function `hill.adapt` allows a data-driven choice of the threshold τ and the estimation of θ_t.

2.3.3 Selection of the bandwidth

We determine the bandwidth h by cross-validation from a sequence of the form $h_l = aq^l$, $l = 0, \ldots, M_h$ with $q = \exp\left(\frac{\ln b - \ln a}{M_h}\right)$, where a is the minimum bandwidth of the sequence, b is the maximum bandwidth of the sequence and M_h is the length of the sequence. The choice is performed globally on the grid $T_{grid} = \{t_1, \ldots, t_K\}$ of points $t_i \in [0, T_{\max}]$, where the number K of the points on the grid is defined by the user. The choice $K = n$ is possible but can be time consuming for large samples. We recommend to use a fraction of n.

We choose h_{cv} by minimizing in h_m, $m = 1, \ldots, M_h$ the cross-validation function

$$CV(h_m, p_{cv}) = \frac{1}{M_h \text{card}(T_{grid})} \sum_{h_l} \sum_{t_i \in T_{grid}} \left| \ln \frac{\widehat{q}^{(-i)}_{p_{cv}}(t_i, h_m)}{\widehat{F}^{-1}_{t_i,h_l}(p_{cv})} \right|, \tag{2.3.9}$$

where $\widehat{F}^{-1}_{t_i,h_l}(p_{cv})$ is the empirical quantile from the observations in the window $[t_i - h_l, t_i + h_l]$, $\widehat{q}^{(-i)}_{p_{cv}}(t_i, h_m)$ is the quantile estimator inside the window $[t_i - h_m, t_i +$

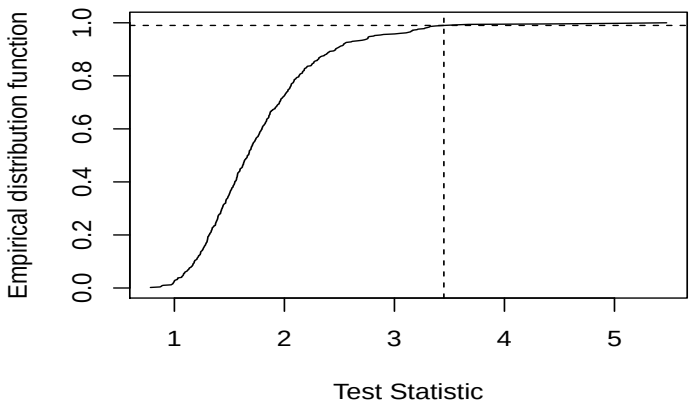

Figure 2.1: Empirical distribution function of the test statistic for the truncated Gaussian kernel with $N_{MC} = 1000$ Monte-Carlo samples of size $n = 500$. The vertical dashed line represents the critical value ($D = 3.4$) corresponding to the 0.99-empirical quantile of the test statistic.

h_m] defined by (2.3.8) with the observation X_{t_i} removed and τ being the adaptive threshold given by the remaining observations inside the window $[t_i - h_m, t_i + h_m]$. The function `bandwidth.CV` selects the bandwidth h by cross-validation.

2.4 Package presentation on simulation study

The **extremefit** package is written in R, ([57]). In this section, we demonstrate the package using its application on two simulated data sets.

The following code displays the computation of the survival probabilities and quantiles using the adaptive choice of the threshold provided by the `hill.adapt` function.

```
R> library("extremefit")
R> set.seed(5)
R> X <- abs(rcauchy(200))
R> n <- 100
R> HA <- hill.adapt(X)
R> predict(HA, newdata = c(3, 5, 7), type = "survival")$p
R> predict(HA, newdata = c(0.9, 0.99, 0.999, 0.9999), type = "quantile")$y
```

A simple use of the method described in Section 2.3 is given by the following

example. With $t_i = i/n$, we consider data $X_{t_1}, \ldots, X_{t_n}$ generated by the Pareto change-point model defined by

$$F_t(x) = \left(1 - x^{-1/2\theta_t}\right) \mathbb{1}_{x \leq \tau} + \left(1 - x^{-1/\theta_t} \tau^{1/2\theta_t}\right) \mathbb{1}_{x > \tau}, \tag{2.4.1}$$

where θ_t is a time varying parameter depending on $t \in [0, 1]$ defined by $\theta_t = 0.5 + 0.25 \sin(2\pi t)$ and $\tau = 3$ as described in Durrieu et al. [54]. We consider the sample size $n = 50000$. The following commands generate one sample from the model (2.4.1).

```
R> library("extremefit")
R> set.seed(5)
R> n <- 50000 ; tau <- 3
R> theta <- function(t){0.5 + 0.25 * sin(2 * pi * t)}
R> T <- 1:n / n; Theta <- theta(T)
R> X <- rparetoCP(n, a0 = 1 / (Theta * 2),
+    a1 = 1 / Theta, x1 = tau)
```

The **extremefit** package provides the estimates of θ_t, $q_p(t, h)$ and $F_t(x)$ for large values of x particularly. We select the bandwidth h_{cv} by minimizing the cross-validation function implemented in `bandwidth.CV`. The weights are computed using the truncated Gaussian kernel ($\sigma = 1$), which is implemented in `TruncGauss.kernel`. This kernel implies $D = 3.4$. To select the bandwidth h_{cv}, we define a grid of possible values of h as indicated in Section 2.3.3 with $a = 0.005$, $b = 0.05$ and $M_h = 20$. Moreover, we fix the parameter $p_{cv} = 0.99$. The parameter T_{grid} define a grid of $t \in T_{grid}$ to perform the cross-validation.

```
R> a <- 0.005 ; b <- 0.05 ; Mh <- 20
R> hl <- bandwidth.grid(a, b, Mh, type = "geometric")
R> Tgrid <- seq(0, 1, 0.02)
R> Hcv <- bandwidth.CV(X, T, Tgrid, hl, pcv = 0.99,
+    kernel = TruncGauss.kernel, CritVal = 3.4, plot = FALSE)
R> Hcv$h.cv
```

```
[1] 0.02727797
```

For each $t \in T_{grid}$, we determine the data-driven threshold τ and the estimates $\widehat{\theta}_{t,h_{cv},\tau}$ using the function `hill.ts`.

```
R> Tgrid <- seq(0, 1, 0.01)
R> hillTs <- hill.ts(X, T, Tgrid, h = Hcv$h.cv,
+    kernel = TruncGauss.kernel, CritVal = 3.4)
```

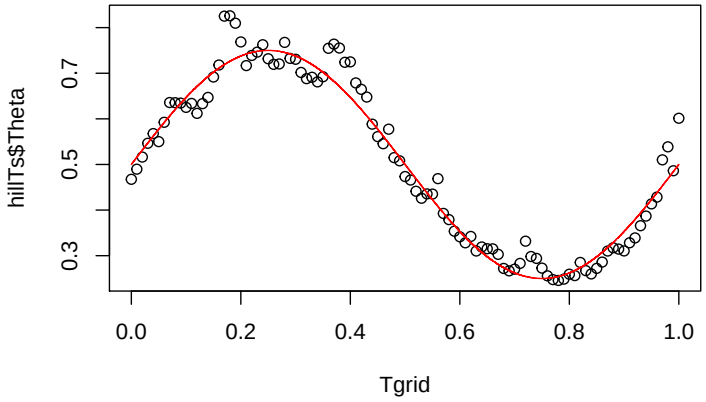

Figure 2.2: Plot of the θ_t estimate $\widehat{\theta}_{t,h_{cv},\tau}$ (black dots) and the true θ_t (red line) for each $t \in T_{grid}$.

For each $t \in T_{grid}$, we display $\widehat{\theta}_{t,h_{cv},\tau}$ and the true value $\theta_t = 0.5 + 0.25\sin(2\pi t)$ in Figure 2.2.

```
R> plot(Tgrid, hillTs$Theta)
R> lines(T, Theta, col = "red")
```

The estimates of the quantiles and the survival probabilities are determined using the command `predict.hill.ts`. For instance the estimate of the p-quantile $F_t^{-1}(p)$ of order $p = 0.99$ and $p = 0.999$ are computed with the following R command:

```
R> p <- c(0.99, 0.999)
R> PredQuant <- predict(hillTs, newdata = p, type = "quantile")
```

Figure 2.3 displays the true and the estimated quantiles of order $p = 0.99$ and $p = 0.999$ of the Pareto change-point distribution defined by (2.4.1). The true quantiles can be accessed by the `qparetoCP` function.

```
R> TrueQuant <- matrix(0, ncol = 2, nrow = n)
R> for(i in 1:length(p)){
+    TrueQuant[,i] <- qparetoCP(p[i], a0 = 1 / (Theta * 2),
+      a1 = 1 / Theta, x1 = tau)
+  }
R> plot(Tgrid, log(as.numeric(PredQuant$y[1,])), ylim = c(1.5, 6.3),
```

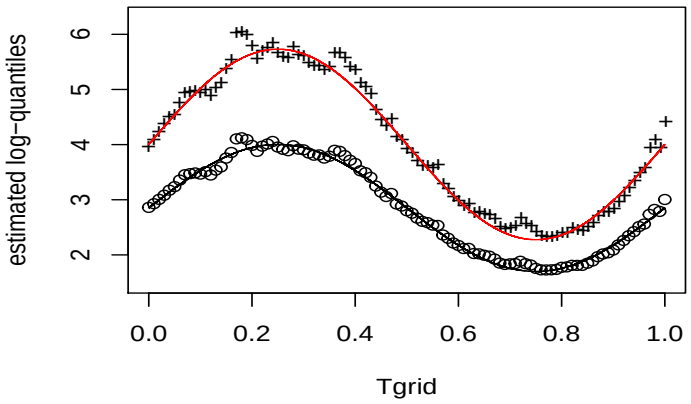

Figure 2.3: Plot of the true log quantiles $F_t^{-1}(p)$ with $p = 0.99$ (black line) and $p = 0.999$ (red line) and the corresponding estimated quantiles with $p = 0.99$ (black dots) and $p = 0.999$ (black cross) as function of $t \in T_{grid}$.

```
+    ylab = "estimated log-quantiles")
R> points(Tgrid, log(as.numeric(PredQuant$y[2,])), pch = "+")
R> lines(T, log(TrueQuant[,1]))
R> lines(T, log(TrueQuant[,2]), col = "red")
```

In the same way, we estimate for each $t \in T_{grid}$ the survival probabilities $S_t(x) = 1 - F_t(x)$. The function `predict.hill.ts` with the option `type = "survival"` computes the estimated survival function for a given x. For each $t \in T_{grid}$, the following commands compute the estimate of $S_t(x)$ for $x = 20$ and $x = 30$.

```
R> x <- c(20, 30)
R> PredSurv <- predict(hillTs, newdata = x, type = "survival")
R> TrueSurv <- matrix(0, ncol = 2, nrow = n)
R> for(i in 1:length(x)){
+    TrueSurv[, i] <- 1 - pparetoCP(x[i], a0 = 1 / (Theta * 2),
+      a1 = 1 / Theta, x1 = tau)
+  }
R> plot(Tgrid, as.numeric(PredSurv$p[1,]),
+    ylab = "estimated survival probabilities")
R> points(Tgrid, as.numeric(PredSurv$p[2,]), pch = "+")
R> lines(T, TrueSurv[,1])
R> lines(T, TrueSurv[,2], col = "red")
```

The estimations of the quantile and the survival function displayed in Figures 2.3 and 2.4 have been performed for one sample. Now we analyze the performance of the

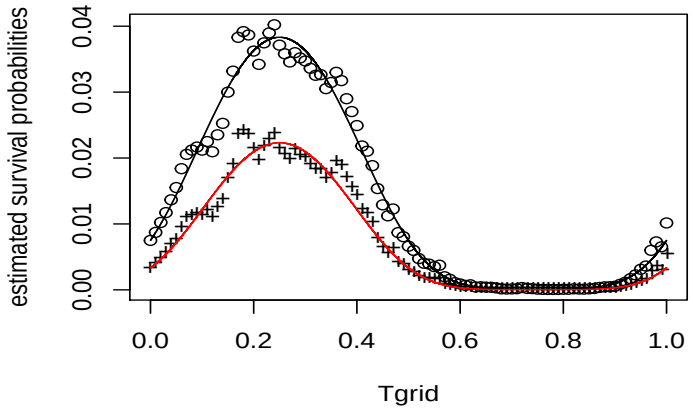

Figure 2.4: Plot of the true survival probabilities $S_t(x)$ at $x = 20$ (black line) and $x = 30$ (red line) and the corresponding estimated survival probabilities at $x = 20$ (black dots) and $x = 30$ (black cross) as function of $t \in T_{grid}$.

quantile estimator via a Monte-Carlo study. We obtain in Figure 2.5 satisfying results on a simulation study using $N_{MC} = 1000$ Monte-Carlo replicates. The iteration of the previous R commands on $N_{MC} = 1000$ simulations are not given because of the computation time.

2.5 Real-world data sets

2.5.1 Wind data

The wind is studied in the framework of an alternative to fossil energy. From wind energy database we focus on wind speed problems. Studies of wind speed in extreme value theory were made to model windstorm losses or detect areas which can be subject to hurricanes (see Rootzén and Tajvidi [17] and Simiu and Heckert [69]).

The wind data in the package **extremefit** comes from Airport of Brest (France) and represents the average wind speed per day from 1976 to 2005. The data set is included in the package **extremefit** and can be loaded by the following code.

```
R> library("extremefit")
R> data("dataWind")
R> attach(dataWind)
```

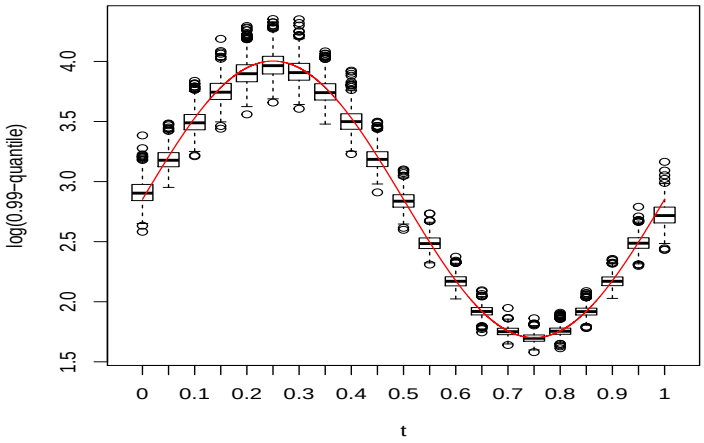

Figure 2.5: Boxplots of the log $\widehat{q}_{0.99}(t, h_{cv})$ adaptive estimators with bandwidth h_{cv} chosen by cross-validation and $t \in T_{grid}$. The true log 0.99-quantile is plotted as a red line.

The commands below illustrate the function `hill.adapt` on the wind data set and computes a monthly estimation of the survival probabilities $1 - \widehat{F}_{t,h,\tau}(x)$ for a given $x = 100$ km/h with the function `predict.hill.adapt`.

```
R> pred <- NULL
R> for(m in 1:12){
+    indices <- which(Month == m)
+    X <- Speed[indices] * 60 * 60 / 1000
+    H <- hill.adapt(X)
+    pred[m] <- predict(H, newdata = 100, type = "survival")$p
+  }
R> plot(pred, ylab = "Estimated survival probability", xlab = "Month")
```

2.5.2 Sea shores water quality

The study of the pollution in the aquatic environment is an important problem. Humans tend to pollute the environment through their activities and the water quality survey is necessary. The bivalve's activity is investigated by modeling the valve movements using high frequency valvometry. The electronic equipment is described in Tran et al. [70] and modified by Chambon et al. [71]. More information can be found in http://molluscan-eye.epoc.u-bordeaux1.fr/.

High-frequency data (10 Hz) are produced by noninvasive valvometric techniques and the study of the bivalve's behavior in their natural habitat leads to the proposal

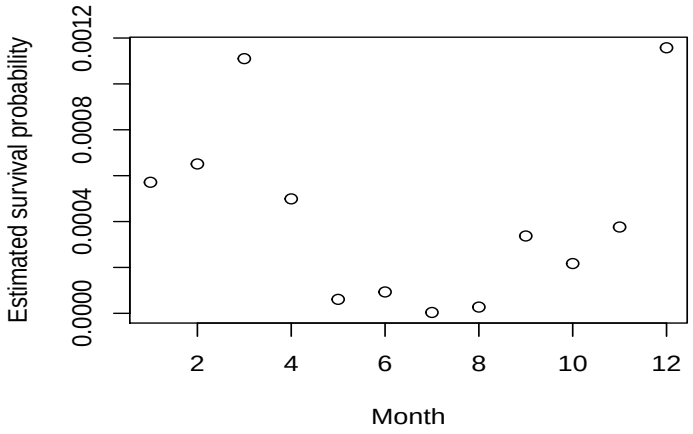

Figure 2.6: Plot of the estimated survival probability $1 - \widehat{F}_{t,h,\tau}(x)$ at $x = 100$ km/h.

of several statistical models (Sow et al. [72], Schmitt et al. [73], Jou and Liao [74], Coudret et al. [75], Azaïs et al. [76], Durrieu et al. [54] and Durrieu et al. [77]). It is observed that in the presence of a pollutant, the activity of the bivalves is modified and consequently they can be used as bioindicators to detect perturbations in aquatic systems (pollutions, global warming). A group of oysters *Crassostrea gigas* of the same age are installed around the world but we concentrate on the Locmariaquer site (GPS coordinates 47°34 N, 2°56 W) in France. The oysters are placed in a traditional oyster bag. In the package **extremefit**, we provide a sample of the measurements for one oyster over one day. The data can be accessed by

```
R> library("extremefit")
R> data("dataOyster")
```

The description of the data can be found with the R command `help(dataOyster)`. The following code covers the velocities and the time covariate and also displays the data.

```
R> Velocity <- dataOyster$data[, 3]
R> time <- dataOyster$data[, 1]
R> plot(time, Velocity, type = "l", xlab = "time (hour)",
+    ylab = "Velocity (mm/s)")
```

We observe in Figure 2.7 that the velocity is equal to zero in two periods of time. To facilitate the study of these data, we have included a time grid where the intervals with null velocities are removed. The grid of time can be accessed by `dataOyster$Tgrid`. We shift the data to be positive.

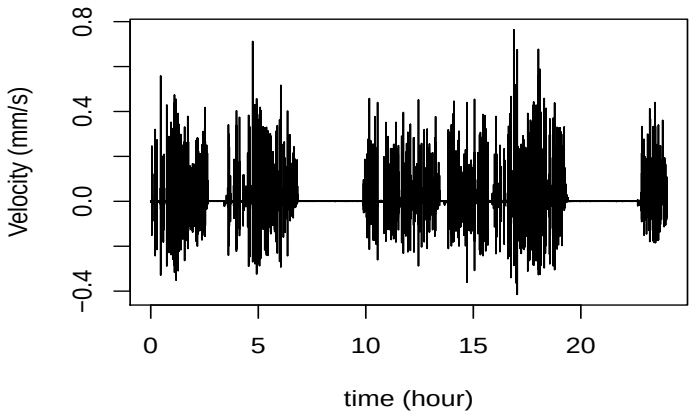

Figure 2.7: Plot of the velocity of the valve closing and opening over one day.

```
R> new.Tgrid <- dataOyster$Tgrid
R> X <- Velocity + (-min(Velocity))
```

The bandwidth parameter is selected by cross-validation method ($h_{cv} = 0.2981812$) using `bandwidth.CV` but we select it manually in the following command due to long computation time.

```
R> hcv <- 0.2981812
R> TS.Oyster <- hill.ts(X ,t = time, new.Tgrid, h = hcv,
+    TruncGauss.kernel, CritVal = 3.4)
```

The estimations of the extreme quantile of order 0.999 and the probabilities of rare events are computed as described in Section 2.3.2. The critical value of the sequential test is $D = 3.4$ when considering a truncated Gaussian kernel, see Table 2.1. A global study on a set of 16 oysters on a 6 months period is given in Durrieu et al. [54].

```
R> pred.quant.Oyster <- predict(TS.Oyster, newdata = 0.999, type = "quantile")
R> plot(time, Velocity, type = "l", ylim = c(-0.5, 1),
+    xlab ="Time (hour)", ylab = "Velocity (mm/s)")
R> quant0.999 <- rep(0, length(seq(0, 24, 0.05)))
R> quant0.999[match(new.Tgrid, seq(0, 24, 0.05))] <-
+    as.numeric(pred.quant.Oyster$y) -
+    (-min(Velocity))
R> lines(seq(0, 24, 0.05), quant0.999, col = "red")
```

In Durrieu et al. [54] and Durrieu et al. [77], we observe that valvometry using extreme value theory allows in real-time *in situ* analysis of the bivalves behavior and appears as an effective early warning tool in ecological risk assessment and marine environment monitoring.

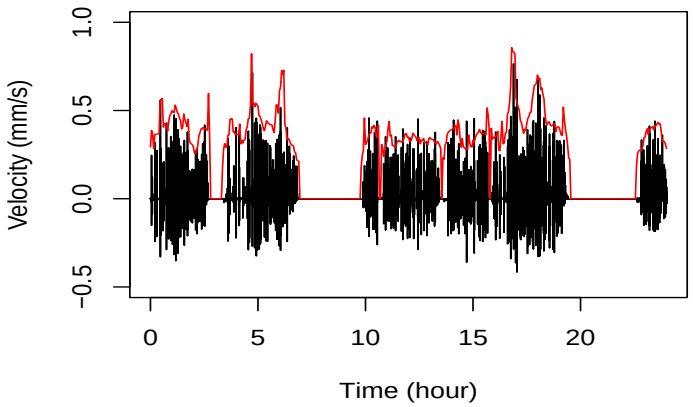

Figure 2.8: Plot of the estimated 0.999-quantile (red line) and the velocities of valve closing (black lines).

2.5.3 Electric consumption

Electric consumption is an important challenge due to population expansion and increasing needs in a lot of countries. Obviously this consumption is dealing with the state procurement policies. Therefore statistical models may give understanding of electric consumption. Multiple models have been used to forecast the electric consumption, as regression and time series models (Bianco et al. [78] and Ranjan and Jain [79], Harris and Liu [80] and Bercu and Proïa [81]). Durand et al. [32] used hidden Markov model to forecast the electric consumption. A research project conducted in France (Lorient, GPS coordinates 47°45 N, 3°22 W) concerns the measurements of electric consumption using Linky, a smart communicating electric meter (http://www.enedis.fr/linky-communicating-meter). Installed in end-consumer's dwellings and linked to a supervision center, this meter is in constant interaction with the network. The Linky electric meter allows a measurement of the electric consumption every 10 minutes.

To prevent from major power outages, the SOLENN project is testing an alternative to load shedding (http://www.smartgrid-solenn.fr/en/). Data of electric consumption are collected on selected dwellings to study the effect of a decrease on the maximal power limit. For example, an dwelling with a maximal electric power

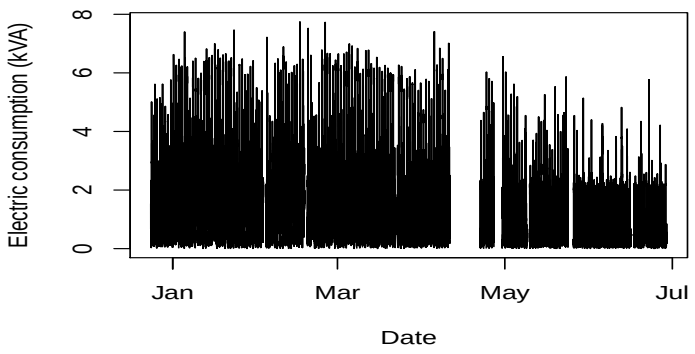

Figure 2.9: Electric consumption of one customer from the 24th December 2015 to the 29th June 2016.

contract of 9 kiloVolt ampere is decreased to 6 kiloVolt ampere. This experiment enables to study the consumption of the dwelling with the application of an electric constraint related to the need of the network. For instance, after an incident such as a power outage on the electric network, the objective is to limit the number of dwellings without electricity. If during the time period where the electric constraint is applied, the electric consumption of the dwelling exceeds the restricted maximal power, the breaker cuts off and the dwelling has no more electricity. The consumer can, at that time, close the circuit breaker and gets the electricity back. In any cases, at the end of the electric constraint, the network manager can close the breaker using the Linky electric meter which is connected to the network. The control of the cut off breakers is crucial to prevent a dissatisfaction from the customers and to detect which dwellings are at risk.

The extreme value modeling approach described in Section 2.3 was carried out on the electric consumption data for one dwelling from the 24th December 2015 to the 29th June 2016. This data are accessible on the **extremefit** package and Figure 2.9 displays the electric consumption of one dwelling. This dwelling has a maximal power contract of 9 kVA.

```
R> data("LoadCurve")
R> Date <- as.POSIXct(LoadCurve$data$Time * 86400, origin = "1970-01-01")
R> plot(Date, LoadCurve$data$Value / 1000, type = "l",
+    ylab = "Electric consumption (kVA)", xlab = "Date")
```

We consider the following grid of time T_{grid}:

```
R> Tgrid <-  seq(min(LoadCurve$data$Time), max(LoadCurve$data$Time),
+    length = 400)
```

We observe in April 2016 missing values in Figure 2.9 due to a technical problem. We modify the grid of time by removing the intervals of T_{grid} with no data.

```
R> new.Tgrid <-LoadCurve$Tgrid
```

We choose the truncated Gaussian kernel and the associated critical value of the goodness-of-fit test is $D = 3.4$ (see Table 2.1). The bandwidth parameter is selected by cross-validation method ($h_{cv} = 3.44$) using `bandwidth.CV` but we select it manually in the following command due to long computation time.

```
R> HH <- hill.ts(LoadCurve$data$Value, LoadCurve$data$Time, new.Tgrid,
+    h = 3.44, kernel = TruncGauss.kernel, CritVal = 3.4)
```

To detect the probability to cut off the breaker, we compute for each time in the grid the estimates of the probability to exceed the maximal power of 9kVA and of the extreme 0.99-quantile. Figure 2.10 displays the electric consumption during the period of study and the estimated quantile of order 0.99.

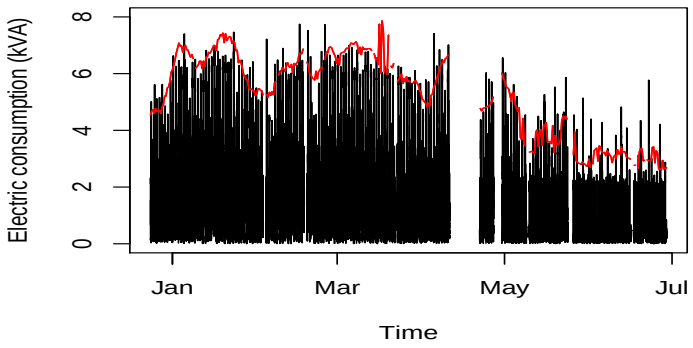

Figure 2.10: Plot of the estimated 0.99-quantile (red line) for each time from the 24th December 2015 to the 29th June 2016.

```
R> Quant <- rep(NA, length(Tgrid))
R> Quant[match(new.Tgrid,Tgrid)] <- as.numeric(predict(HH,
+    newdata = 0.99, type = "quantile")$y)
R> plot(Date, LoadCurve$data$Value/1000, ylim = c(0, 8),
+    type = "l", ylab = "Electric consumption (kVA)", xlab = "Time")
R> lines(as.POSIXlt((Tgrid) * 86400, origin = "1970-01-01",
+    tz = "Europe/Paris"), Quant / 1000, col = "red")
```

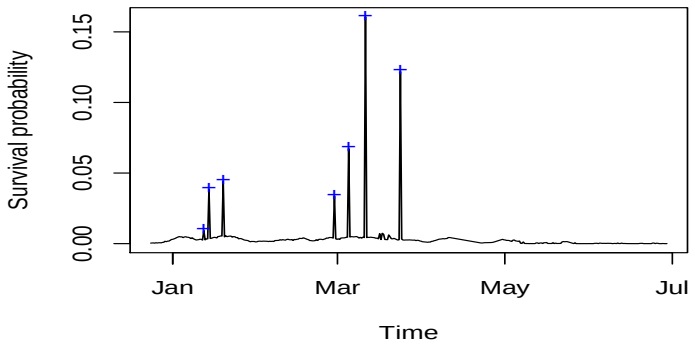

Figure 2.11: Plot of the estimated survival probability depending on the time and the maximal power. The plus symbol correspond to the electric constraint period.

The plus symbol appear at the times when the maximal power was decreased corresponding to the 11th, 13th, 18th of January, the 25th of February and the 1st, 7th, 18th of March in 2016, with a respectively decrease to 6.3, 4.5, 4.5, 4.5, 3.6, 2.7 and 2.7 kVA. Figure 2.11 displays the estimated survival probability which depends on the time and the maximal power. We can observe that the survival probability is higher than usual during this period. Furthermore, we have the auxiliary information that this dwelling cuts off its breaker on every electric constraint period except the 11th of January, 2016.

Using prediction method coupled with the method implemented in the package **extremefit**, it will be possible to detect high probability of cutting off a breaker and react accordingly.

2.6 Conclusion

This paper focus on the functions contained in the package **extremefit** to estimate extreme quantiles and probabilities of rare events. The package **extremefit** works also well on large data set and the performance was illustrated on simulated data and on real data sets.

The choice of the pointwise data driven threshold allows a flexible estimation of the extreme quantiles. The diffusion of the use of this method for the scientific community will improve the choice of estimation of the extreme quantiles and probability of rare events using the peak-over-threshold approach.

Chapter 3

Estimation of extreme survival probabilities with Cox model

This chapter is the subject of an article in collaboration with Ion Grama which is waiting for a second reply from the reviewers.

RÉSUMÉ. Nous proposons une extension au modèle de Cox à hasards proportionnels qui permet une estimation des probabilités d'événements rare. Lorsque les données sont soumises à un haut taux de censure, l'estimation de la queue de la distribution de survie n'est pas fiable. Pour améliorer l'estimation de la fonction de survie "baseline" en dehors du rayon des données observées, nous ajustons un modèle des valeurs extrêmes à la queue de la distribution "baseline" au-delà d'un certain seuil sous certaines conditions. Les fonctions de survies conditionnelles aux covariables sont facilement calculées à partir de la fonction "baseline". Deux procédures proposant un choix automatique du seuil et une estimation agrégée des probabilités de survie sont proposés. La performance du modèle est étudié par simulation et une application sur deux jeux de données réelles est donné.

ABSTRACT. We propose an extension of the regular Cox's proportional hazards model which allows the estimation of the probabilities of rare events. It is known that when the data are heavily censored, the estimation of the tail of the survival distribution is not reliable. To improve the estimate of the baseline survival function in the range of the largest observed data and to extend it outside, we adjust the tail of the baseline distribution beyond some threshold by an extreme value model under appropriate assumptions. The survival distributions conditioned to the covariates are easily computed from the baseline. A procedure allowing an automatic choice of the threshold and an aggregated estimate of the survival probabilities are also proposed. The performance is studied by simulations and an application on two data set is given.

3.1 Introduction

The proportional hazards model introduced by [6] has been largely studied over the years and multiple extensions have been made to the original model. These developments often aim to make inference on the regression parameters in various setting including censoring, time-dependent coefficients, stratified and multistates models, missing and incomplete data etc. The references [30], [82], [26] and [25] give an overall view of this subject. The estimation of the underlying hazard functions is an important ingredient which mostly follows the interrogation of knowing the effect of a treatment. We refer to [83], [84] among others for an illustration. If we are in presence of a significant amount of censored data it is well known that we cannot predict with sufficient precision how the tail of the estimated survival distribution behaves near or beyond the last observed value. This is illustrated in the Figure 3.1 (top) using a real data example, where we can observe that the estimated baseline survival probability becomes constant in the long run.

The present paper aims to estimate the values of the survival distributions conditionally to a covariate in the Cox's proportional hazards model in the case when the estimated probabilities are out of the range of the observed data by using extreme value modeling. Our analysis is based on the Peak-Over-Threshold method (see [14]), which allows to estimate the tail of a distribution beyond a threshold. Applications of this method can be seen in various domain such as insurance, biology, ecology, whether forecast etc. For insurance and financial applications, we can refer to [16] and [12] among many others. A rainfall data study can be found in [11] and high-frequency oyster data are studied in [54]. For some results related to the estimation of extreme values under random censoring without covariates we refer to [85], [86] and [20]. The case with covariates has been considered in [21], [87] and [88]. However, to the best of our knowledge, this approach was not used so far in the context of the Cox model with covariates.

Our idea is simple: we adjust a Pareto distribution for observations beyond a threshold, while the remaining part is estimated nonparametrically by the Nelson-Aalen estimator. The main difficulty is the appropriate choice of the threshold, which can be problematic as a large value will lead to an important variability and small value will increase the bias. This quandary is well known in theory of extreme values. To choose the threshold we make use of the adaptive procedure based on consecutive tests developed in [51]. In addition to this, we propose an aggregation procedure which allows to improve significantly the stability of the estimation. The performance of the proposed estimators is demonstrated by a simulation study and some applications are given. As an illustration in Figure 3.1 (bottom) we show the estimated baseline survival function by the proposed method.

The paper [89] deals also with the Cox model but in a very different way. In [89] we dealt only with the Cox model with qualitative covariates, say z_i with a

finite number of modalities. Each modality has its own choice of the threshold $\widehat{\tau}$, which in principle are different. In the present paper we estimate only one threshold $\widehat{\tau}$ under the baseline distribution, which is automatically translated to the other modalities of the covariates. Thus we can deal with any type of covariates, and we have only one choice of the threshold. We also note that the paper [54] deals with a non stationary time series setting, where the goal is to estimate high quantiles driven by a non-parametricaly changing distribution function.

The paper is organized as follows. In Section 3.2, we introduce the notations, formulate the model and we state the main results. An explicit computation of the convergence rate using the Hall model is given as an example in Section 3.3. An automatic selection procedure of the threshold is stated in Section 3.4. In Section 3.5 we formulate our procedure for the aggregation of the estimated survival probabilities. A simulation study is done in Section 3.7 and an application on two data sets is given in Section 3.8.

3.2 Main results

3.2.1 Notations and model

Denote by X a random variable representing the failure time, by C a random variable representing the censoring time and by Z a random covariate vector. We assume that X and C admit positive densities on $[x_0, \infty)$, with $x_0 \geq 0$ and that X and C are independent conditionally to Z. The observation time and failure indicator are respectively

$$T = \min\{X, C\} \quad \text{and} \quad \Delta = \mathbb{1}_{X \leq C},$$

where $\mathbb{1}$ is the indicator function. The Cox model (introduced by [6]) specifies that the hazard function of the failure time X depends on the value z of covariate vector Z as follows:

$$h(x \,|\, z) = \exp(\beta \cdot z) h_0(x), \;\; x \geq x_0,$$

where $x_0 \geq 0$, β is a vector of parameters, h_0 is an unknown baseline hazard function and $\beta \cdot z$ denotes the scalar product between β and z. We denote by $f(x \,|\, z)$ and $S(x \,|\, z) = 1 - F(x \,|\, z)$, respectively the density and survival functions of the failure time X given $Z = z$. The hazard function $h(x \,|\, z)$ is related to the functions $f(x \,|\, z)$ and $S(x \,|\, z)$ by the expressions

$$h(x \,|\, z) = f(x \,|\, z)/S(x \,|\, z)$$

and

$$S(x \,|\, z) = \exp(-\int_{x_0}^{x} h(u \,|\, z) du).$$

Similarly, the hazard function of the censoring time is denoted by $h_C(\cdot)$, the density and survival function are respectively denoted $f_C(\cdot)$ and $S_C(\cdot) = 1 - F_C(\cdot)$. Let

$$S_0(x) = S(x \,|\, 0) = \exp(-\int_{x_0}^{x} h_0(u)du) \tag{3.2.1}$$

be the baseline survival function. The survival function $S(\cdot \,|\, z)$ is related to the baseline survival function $S_0(\cdot)$ by the expression

$$S(x \,|\, z) = S_0(x)^{\exp(\beta \cdot z)}.$$

Assume that we observe a sequence of independent triples (t_i, δ_i, z_i), $i = 1, ..., n$, where all z_i's are nonrandom and each pair (t_i, δ_i) has the law of (T, Δ) given $Z = z_i$. In this paper we address the question of estimating the survival function $S(x \,|\, z)$ for large values of x.

Let us explain the difficulties related to this problem using the classical Nelson-Aalen estimator. In the case when x is larger than the last observed time $t_{\max} = \max\{t_1, ..., t_n\}$, the Nelson-Aalen estimator $\widehat{S}_{NA}(\cdot \,|\, z)$ of $S(\cdot \,|\, z)$ takes two positive constant values depending on the fact that the last observed time is censored or not. In Figure 3.1 (top) we plot the estimated baseline survival function $\widehat{S}_{NA}(\cdot \,|\, 0)$ for the commonly accessible `bladder` data set from `R` package `survival` (see Section 3.8.1 for details). Note that the Nelson-Aalen curve $\widehat{S}_{NA}(t \,|\, 0)$ becomes constant for $t \geq t_{\max}$. Moreover, as the survival times are heavily censored, the estimated survival probability $\widehat{S}_{NA}(t \,|\, 0)$ is far above 0 for all $t \geq t_{\max}$. To overcome this effect, we assume some additional constraints on the survival function, which allow us to extrapolate it outside the available data range. Specifically, we assume the following condition:

C1. We assume that F_0 belongs to the maximal domain of attraction of the Fréchet law with extreme value index $\theta > 0$ which means that there exists two sequences $a_n > 0$ and b_n such that, for any $x \geq 0$,

$$F_0^n(a_n x + b_n) \to \Phi_\theta(x) \text{ as } n \to \infty,$$

where $\Phi_\theta(x) = e^{-x^{-\theta}}$, $x \geq 0$ is the Fréchet law and $F_0^n(x) = \mathbb{P}(\max_{1 \leq i \leq n} x_i \leq x)$, with $x_1, x_2, \ldots$ an i.i.d. sequence of random variables of common distribution F_0.

By the Fisher-Tippett-Gnedenko theorem (see Theorem 2.1 page 75 in [5]), condition (**C1**) is equivalent to the property that for each $x \geq 1$,

$$\frac{S_0(\tau x)}{S_0(\tau)} \to (x)^{-1/\theta} \text{ as } \tau \to \infty. \tag{3.2.2}$$

As a consequence of (3.2.2) the following semi-parametric model is considered for the baseline survival function:

$$S_{0,\tau,\theta}(x) = \begin{cases} S_0(x) & \text{if } x \in [0, \tau], \\ S_0(\tau) \left(\frac{x}{\tau}\right)^{-1/\theta} & \text{if } x > \tau, \end{cases} \tag{3.2.3}$$

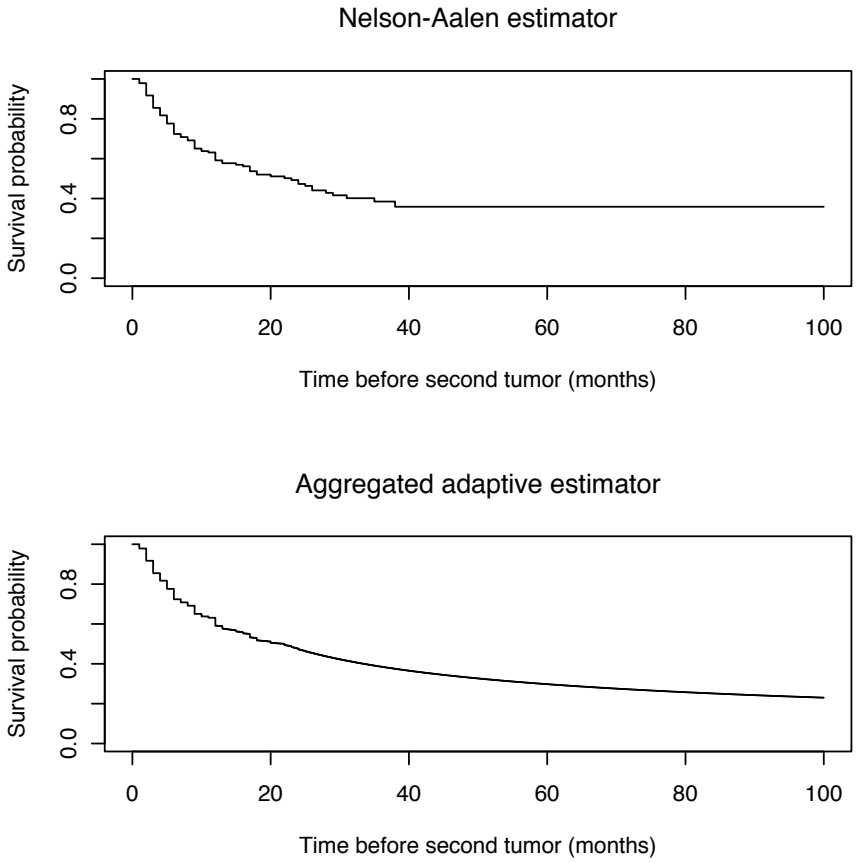

Figure 3.1: Estimated baseline survival probability (no-treatment) of the time of having the first recurrence of the bladder tumor: using the Nelson-Aalen estimator defined by (3.2.9) (top) and the adaptive aggregation defined by (3.5.2) (bottom).

where the function S_0 is fully non-parametric, τ is a threshold parameter and the parametric part is completely described by the Pareto model with parameter θ. We denote the baseline hazard function of the previous model by

$$h_{0,\tau,\theta}(x) = \begin{cases} h_0(x) & \text{if } x \in [0,\tau], \\ \frac{1}{\theta x} & \text{if } x > \tau. \end{cases} \tag{3.2.4}$$

The corresponding cumulative hazard and survival function under the covariate constraint $Z = z$ are then respectively given by $H_{z,\tau,\theta}(x) = e^{\beta \cdot z} \int_{x_0}^{x} h_0(x)$ and

$$S_{z,\tau,\theta}(x) = S(x \mid z, \tau, \theta) = S_{0,\tau,\theta}(x)^{e^{\beta \cdot z}}. \tag{3.2.5}$$

For illustration purposes the estimated baseline survival function S_0 by the proposed model is given in Figure 3.1 (bottom), where for the estimation we have used the aggregation procedure with the adaptive choice of the threshold τ described in Section 3.5.2.

3.2.2 Estimators

In this section, we aim to provide the estimators necessary to deal with the model (3.2.3). To this end, we suppose the regression parameter β of the Cox model known, for example, estimated by the standard procedure described in [6]. As to the threshold τ, it is considered to be fixed (a selection procedure is presented latter on in Section 3.4). To estimate (3.2.3), we will combine the extreme value Hill type estimator of the parameter θ of the tail of the distribution function F_0 and the Nelson-Aalen non-parametric estimator of the baseline survival function S_0.

The joint density of the vector (T, Δ), given $Z = z$, is computed as

$$p_{S_0}(t, \delta \,|\, z) = (e^{\beta \cdot z} h_0(t))^{\delta} S_0(t)^{e^{\beta \cdot z}} f_C(t)^{1-\delta} S_C(t)^{\delta}, \tag{3.2.6}$$

where $t \in [x_0, \infty)$ and $\delta = \{0, 1\}$. Denote by $p_{S_{0,\tau,\theta}}(t, \delta \,|\, z)$ the joint density of the vector (T, Δ), given $Z = z$, when the survival function $S(t \,|\, z)$ obeys the model (3.2.5):

$$p_{S_{0,\tau,\theta}}(t, \delta \,|\, z) = (e^{\beta \cdot z} h_{0,\tau,\theta}(t))^{\delta} S_{0,\tau,\theta}(t)^{e^{\beta \cdot z}} f_C(t)^{1-\delta} S_C(t)^{\delta}.$$

Then the quasi-log-likelihood of the model is

$$\mathcal{L}(\theta \,|\, \mathbf{z}) = \mathcal{L}(\theta \,|\, z_1, ..., z_n) = \sum_{i=1}^{n} \ln p_{S_{0,\tau,\theta}}(t_i, \delta_i \,|\, z_i).$$

Removing the terms related to the censoring, the partial quasi-log-likelihood is

$$\mathcal{L}^{part}(\theta \,|\, \mathbf{z}) = \sum_{i=1}^{n} \left(\delta_i \ln(h_{0,\tau,\theta}(t_i)) + (\beta \cdot z_i)\delta_i - e^{\beta \cdot z_i} \int_{x_0}^{t_i} h_{0,\tau,\theta}(u) du \right),$$

where the baseline hazard function $h_{0,\tau,\theta}(.)$ is defined by (3.2.4) and $\mathbf{z} = z_1, \cdots, z_n$. Now, maximizing $\mathcal{L}^{part}(\theta \,|\, \mathbf{z})$ with respect to θ, yields the estimator

$$\widehat{\theta}_{\tau} = \frac{\sum_{t_i > \tau} e^{\beta \cdot z_i} \ln\left(\frac{t_i}{\tau}\right)}{\sum_{t_i > \tau} \delta_i}. \tag{3.2.7}$$

One can see that, the estimator is a transformation of the estimator introduced in [90].

The Nelson-Aalen estimator, suggested by [7] and rediscovered by [8], focus on estimating the cumulative hazard function $H_0(x) = \int_0^x h_0(t)dt$ by

$$\widehat{H_0}(t) = \sum_{t_i \leq t} \widehat{h}_0(t_i),$$

where, by maximizing the partial quasi-log-likelihood, we have

$$\widehat{h}_0(t_i) = \frac{\delta_i}{\sum_{j=1}^{n} e^{\beta \cdot z_j} \mathbb{1}_{t_j \geq t_i}}.$$

This estimator was suggested by [29]. The estimator of the survival function (3.2.3) is then given by

$$\widehat{S}_{0,\tau,\widehat{\theta}_\tau}(t) = \begin{cases} \widehat{S}_0(t) & \text{if } t \in [0,\tau], \\ \widehat{S}_0(\tau)\left(\frac{t}{\tau}\right)^{-1/\widehat{\theta}_\tau} & \text{if } t > \tau, \end{cases} \tag{3.2.8}$$

where the Nelson-Aalen non-parametric estimator of S_0 is defined by

$$\widehat{S_0}(t) = e^{-\widehat{H}_0(t)}. \tag{3.2.9}$$

The estimator of the survival function (3.2.5) is then given by $\widehat{S}_{z,\tau,\theta}(t) = \widehat{S}_{0,\tau,\widehat{\theta}_\tau}(t)^{e^{\beta\cdot z}}$.

3.2.3 Consistency of $\widehat{\theta}_\tau$

In this section we state a general consistency result for the estimator $\widehat{\theta}_\tau$ of θ, which we apply in Section 3.3 to obtain the rate of convergence under the Hall model. To state it, we need some notations.

The Kullback-Leibler divergence between two equivalent distributions, say P and Q, is denoted $\mathcal{K}(P,Q) = \int \ln(dP/dQ)dP$. This divergence, between two Pareto distributions with parameters θ' and θ, can be written as

$$\mathcal{K}(\theta',\theta) = \frac{\theta'}{\theta} - 1 - \ln(\frac{\theta'}{\theta}) \sim \left(\frac{\theta'}{\theta} - 1\right)^2 \quad \text{as} \quad \frac{\theta'}{\theta} \to 1. \tag{3.2.10}$$

The χ^2-entropy between the probability measures P and Q is defined by

$$\chi^2(P,Q) = \int dP/dQdP - 1. \tag{3.2.11}$$

The following theorem gives an estimate of the Kullback-Leibler entropy between $\widehat{\theta}_\tau$ and θ which is expressed in terms of the χ^2-entropy between the two laws $P_{S_0}(.\,|\,z_i)$ and $P_{S_{0,\tau,\theta}}(.\,|\,z_i)$. The notation $a_n = O(b_n)$ means that there is a positive constant c such that $\mathbb{P}(a_n > cb_n, b_n < \infty) \to 0$ as $n \to \infty$. In the sequel we denote by $\mathbb{P}$ the probability measure corresponding to the "true" model which has the baseline survival function S_0.

Theorem 3.2.1. *Assume that S_0 is an arbitrary survival function as defined by (3.2.1) and that $S_{0,\tau,\theta}$ satisfies the model (3.2.3). Then for any $\theta > 0$, and $\tau \geq x_0$, we have*

$$\mathcal{K}(\widehat{\theta}_\tau,\theta) = O_{\mathbb{P}}\left(\frac{1}{\widehat{n}_\tau}\sum_{i=1}^{n}\chi^2(P_{S_0}(.\,|\,z_i), P_{S_{0,\tau,\theta}}(.\,|\,z_i)) + \frac{4\ln(n)}{\widehat{n}_\tau}\right),$$

where $\widehat{n}_\tau = \sum_{t_i>\tau}\delta_i$.

Theorem 3.2.1 gives an upper bound of the Kullback-Leibler entropy which can be read in two parts, the bias term $\frac{1}{\widehat{n}_\tau}\sum_{i=1}^{n}\chi^2(P_{S_0}(.\,|\,z_i), P_{S_{0,\tau,\theta}}(.\,|\,z_i))$ and the variance term $\frac{4\ln(n)}{\widehat{n}_\tau}$.

Now, we formulate some sufficient conditions for the consistence of the estimator $\widehat{\theta}_\tau$. In order to estimate the bias term, we need to introduce a quantity which show how the censoring rate evolves with the threshold τ, when we consider only the observations exceeding τ. For this, we introduce the following conditioned mean censoring rate function given $Z = z$:

$$\tau \mapsto q_F(\tau\,|\,z) = \int_\tau^\infty \frac{S(t\,|\,z)}{S(\tau\,|\,z)}\frac{f_C(t\,|\,z)}{S_C(\tau\,|\,z)}dt \in [0,1], \quad \tau \geq x_0.$$

The value $q_F(\tau\,|\,z)$ gives the rate of censored observations above the threshold τ. In order to estimate the extreme survival probabilities it is natural to require that the rate of censored observations above the threshold τ is strictly less than 1. We shall impose on $q_F(\tau\,|\,z)$ the following condition:

C2. There is a constant $q_0 < 1$ such that, for $\tau \geq x_0$ large enough,

$$q_F(\tau\,|\,z_i) \leq q_0, \quad i = 1, ..., n.$$

Condition (**C2**) is easily verified, for instance, when both the distribution function of the survival time and that of the censoring time follow the Cox model and are in the maximal domain of attraction of the Fréchet law with parameters $\theta(z)$ and $\theta_C(z)$ respectively. Indeed, we show in the Lemma 3.B.3 that, in this case, for any z,

$$q_F(\tau\,|\,z) \to \frac{\theta(z)}{\theta(z)+\theta_C(z)} \text{ as } \tau \to \infty.$$

where under some mild assumptions one can verify that

$$\frac{\theta(z)}{\theta(z)+\theta_C(z)} \leq q_0 < 1. \tag{3.2.12}$$

In addition we introduce the following conditions:

C3. There exists $z_{\min}$ and $z_{\max}$ such that

$$z \in \mathbb{Z} \quad := \quad [z_{\min}; z_{\max}].$$

C4. Von-Mises condition:

$$th_0(t) \to \frac{1}{\theta} \quad \text{as} \quad t \to \infty.$$

It is well known that (**C4**) implies condition (**C1**), see [5]. For any $\tau > 0$, set

$$\rho_\tau = \sup_{t>\tau} \left| t h_0(t) - \frac{1}{\theta} \right|.$$

It is easy to see that the Von-Mises condition (**C4**) is equivalent to the fact that there exists a sequence of thresholds (τ_n) satisfying $x_0 \leq \tau_n \to \infty$ as $n \to \infty$ such that $n\rho_{\tau_n}^2 \to 0$ as $n \to \infty$.

The following result shows the consistency of the estimated parameter $\widehat{\theta}_{\tau_n}$:

Theorem 3.2.2. *Assume conditions* (**C2**), (**C3**) *and* (**C4**). *For any sequence* (τ_n) *satisfying*

$$x_0 \leq \tau_n \to \infty, \quad n\rho_{\tau_n}^2 \to 0 \quad as \quad n \to \infty \tag{3.2.13}$$

and

$$\sum_{i=1}^{n} S(\tau_n \,|\, z_i) S_C(\tau_n \,|\, z_i) \to \infty \quad as \quad n \to \infty, \tag{3.2.14}$$

it holds

$$\widehat{\theta}_{\tau_n} \xrightarrow[n\to\infty]{\mathbb{P}} \theta.$$

Condition (3.2.13) gives a control on the bias $\frac{1}{n_\tau} \sum_{i=1}^{n} \chi^2(P_{S_0}(.\,|\,z_i), P_{S_{0,\tau,\theta}}(.\,|\,z_i))$ of the model which should be small as $n \to \infty$, while condition (3.2.14) is responsible for the control of the variance $\frac{4\ln(n)}{\widehat{n}_\tau}$ of the model. We will show in the next section how to verify conditions (**C2**), (3.2.13) and (3.2.14) with a large class of models.

3.3 Computation of the explicit rate of convergence for the Hall model

In this section we consider a model which is related to the families of distributions in [91], [92] and [51] for the extreme value estimation. The result of the Theorem 3.B.4 in Section 3.2.3 shows that the rate of convergence of the estimator $\widehat{\theta}_{\tau_n}$ depends on the threshold τ_n and the survival functions of the survival and censoring times. To express the rate of convergence in terms of the sample size, some assumptions must be made on the survival functions S and S_C.

C5. The baseline hazard function h_0 is such that for some unknown parameter $\theta \in (\theta_{min}, \theta_{max})$ and any $t > 1$,

$$|t h_0(t) - \frac{1}{\theta}| \leq c_1 t^{-\frac{\alpha}{\theta}},$$

where α, c_1, θ_{min} and θ_{max} are some positive constants.

Condition (**C5**) means that $th_0(t)$ converges to $\frac{1}{\theta}$ polynomially fast as $t \to \infty$. Similarly we assume:

C6. The hazard function h_C of the censoring time C is such that for any $z \in \mathbb{Z} := [z_{\min}; z_{\max}]$ and any $t > 1$,

$$|th_C(t \,|\, z) - \frac{1}{\theta_C(z)}| \leq c_2 t^{-\mu},$$

where $\theta_C(z) = \frac{\theta}{\gamma} e^{-\beta \cdot z}$ and $\gamma > 0$, $c_2 > 0$ and $\mu > 1$ are some constants.

Theorem 3.3.1. *Assume condition* (**C3**), (**C5**) *and* (**C6**). *Suppose in addition that* $\frac{\varrho_2}{\varrho_1} \frac{1+\gamma}{1+\gamma+2\alpha} \leq 1$, *where* $\varrho_1 = \min_{\beta \cdot z}(e^{\beta \cdot z})$ *and* $\varrho_2 = \max_{\beta \cdot z}(e^{\beta \cdot z})$. *Then, there exists a constant* c *such that,*

$$\lim_{n \to \infty} \mathbb{P}\left(\mathcal{K}(\widehat{\theta}_{\tau_n}, \theta) \leq c \left(\frac{\ln n}{n} \right)^{1 - \frac{\varrho_2}{\varrho_1} \frac{1+\gamma}{1+\gamma+2\alpha}} \right) = 1,$$

where

$$\tau_n = n^{\frac{\theta/\varrho_1}{1+\gamma+2\alpha}} \ln^{-\frac{\theta/\varrho_1}{1+\gamma+2\alpha}} n.$$

When the covariate z is absent, say $z = 0$, then $\varrho_2 = \varrho_1 = 1$ and we recover the result of Theorem 4.2 of the paper [89], where it is shown that when γ goes to 0 (no censoring) the rate becomes close to the optimal rate of convergence $n^{\frac{2\alpha}{1+2\alpha}}$ in the context of the extreme value estimation, see [93] and [51]. In the case of a binary covariate (i.e. $z \in \{0, 1\}$), if we assume that $\beta > 0$, the convergence speed becomes $\mathcal{K}(\widehat{\theta}_{\tau_n}, \theta) = O_{\mathbb{P}}\left((\frac{\ln n}{n})^{1 - e^{\beta} \frac{1+\gamma}{1+\gamma+2\alpha}} \right)$ with $\tau_n = n^{\frac{\theta}{1+\gamma+2\alpha}} \ln^{-\frac{\theta}{1+\gamma+2\alpha}} n$.

Condition (**C6**) may seem a bit cumbersome at first sight, however, with a closer look we see that it is quite natural if we want to obtain a close rate of convergence. To see this we note that condition (**C5**) can be equivalently stated as

$$|th(t \,|\, z) - \frac{1}{\theta(z)}| \leq c_1 t^{-\frac{\alpha}{\theta(z)}},$$

where the tail index $\theta(z) = \theta e^{-\beta \cdot z}$ depends on the covariate z. Now conditions (3.2.12) and (**C2**) will be verified with $q_0 = \frac{\gamma}{\gamma+1}$.

Condition (**C6**) can be replaced by the following condition:

C7. The hazard function h_C of the censoring time C is such that for any $z \in \mathbb{Z} := [z_{\min}; z_{\max}]$ and any $t > 1$,

$$|th_C(t) - \frac{\gamma}{\theta}| \leq c_2 t^{-\mu},$$

where $\gamma > 0$, $c_2 > 0$ and $\mu > 1$ are some constants.

With (C7) instead of (C6) we can obtain the following result:

Theorem 3.3.2. *Assume condition* (C3), (C5) *and* (C7). *Suppose in addition that* $\frac{\varrho_2+\gamma}{\varrho_1+\gamma+2\alpha\varrho_1} \leq 1$, *where* $\varrho_1 = \min_{\beta\cdot z}(e^{\beta\cdot z})$ *and* $\varrho_2 = \max_{\beta\cdot z}(e^{\beta\cdot z})$. *Then, there exists a constant* c *such that,*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{K}(\widehat{\theta}_{\tau_n}, \theta) \leq c\left(\frac{\ln n}{n}\right)^{1-\frac{\varrho_2+\gamma}{\varrho_1+\gamma+2\alpha\varrho_1}}\right) = 1,$$

where

$$\tau_n = n^{\frac{\theta}{\varrho_1+\gamma+2\alpha\varrho_1}} \ln^{-\frac{\theta}{\varrho_1+\gamma+2\alpha\varrho_1}} n.$$

The proof of Theorem 3.3.2 is similar to that of Theorem 3.3.1 and therefore will not be given in this paper.

3.4 Automatic selection of the threshold

It is well-known that the choice of the threshold τ has a major impact on the quality of the estimation in the extreme value modelling. We propose a data-driven choice of the threshold τ inspired by [51]. The adaptive threshold $\widehat{\tau}$ is selected by a sequential testing procedure followed by a selection using a penalized maximum likelihood.

Consider the following semi-parametric survival function consisting of three parts:

$$S_{0,\tau,\theta,s,\lambda}(x) = \begin{cases} S_0(x) & \text{if } x \in [0, s], \\ S_0(s)\left(\frac{x}{s}\right)^{-\frac{1}{\lambda}} & \text{if } x \in (s, \tau]. \\ S_0(s)\left(\frac{\tau}{s}\right)^{-\frac{1}{\lambda}}\left(\frac{x}{\tau}\right)^{-\frac{1}{\theta}} & \text{if } x > \tau, \end{cases} \tag{3.4.1}$$

where $\lambda > 0$, $\theta > 0$ and $x_0 < s < \tau$. The maximum quasi-likelihood estimators $\widehat{\theta}_\tau$ of θ and $\widehat{\lambda}_{s,\tau}$ of λ are respectively given by (3.2.7) and

$$\widehat{\lambda}_{s,\tau} = \frac{\widehat{\theta}_s \widehat{n}_s - \widehat{\theta}_\tau \widehat{n}_\tau}{\widehat{n}_{s,\tau}},$$

where for brevity, we have denoted $\widehat{n}_s = \sum_{t_i>s} \delta_i$, $\widehat{n}_\tau = \sum_{t_i>\tau} \delta_i$ and $\widehat{n}_{s,\tau} = \sum_{s<t_i<\tau} \delta_i$.

Assume that the observations t_i are ordered in the decreasing order such that $t_1 > ... > t_n$. We define a uniform grid K in the subscripts $i = 1, .., n$ of a size n_{grid}, say $K = \{k_1, k_2, ..., k_{n_{grid}}\}$. The grid K define the set of observations $\{t_{k_1}, t_{k_2}, ..., t_{k_{n_{grid}}}\}$ on which the testing procedure will be performed.

We start with the first subscript $k = k_1$ on the grid K. For the subscript k we test the null hypothesis

$$\mathcal{H}_0(t_k) \quad : \quad S_0(x) = S_{0,t_k,\theta}(x)$$

against the alternative hypothesis

$$\mathcal{H}_1(t_k, t_l) \quad : \quad S_0(x) = S_{0,t_k,\theta,t_l,\lambda}(x),$$

where $S_{0,t_k,\theta}(x)$ is given by (3.2.3), $S_0(x) = S_{0,t_k,\theta,t_l,\lambda}(x)$ is given by (3.4.1), and t_l can change between t_1 and t_k. If we choose all the t_l between t_1 and t_k, some bias is introduced by the observations too close to t_1 and t_k. To overcome this problem, we introduce two parameters ζ' and ζ'' satisfying $0 < \zeta', \zeta'' < 0.5$ which are empirically calibrated. Now t_l will be varying between $t_{(1-\zeta'')k}$ and $t_{\zeta' k}$.

The log-likelihood ratio test statistic used to test the null hypothesis $\mathcal{H}_0(t_k)$ against the alternative $\mathcal{H}_1(t_k, t_l)$ is given by

$$LR(t_k, t_l) = \widehat{n}_{t_k,t_l} \mathcal{K}(\widehat{\lambda}_{t_k,t_l}, \widehat{\theta}_{t_k}) + \widehat{n}_{t_l} \mathcal{K}(\widehat{\theta}_{t_l}, \widehat{\theta}_{t_k}), \tag{3.4.2}$$

where $\mathcal{K}$ is the Kullback-Leibler divergence defined in Section 3.2.3. The test statistic is compared to a critical value D, which is also empirically calibrated. We test, for every $t_l \in [t_{(1-\zeta'')k}; t_{\zeta' k}]$, the hypothesis $\mathcal{H}_0(t_k)$ against the alternative $\mathcal{H}_1(t_k, t_l)$ and, if the critical value is not exceeded, we increase the subscript on the grid K and preform the test with the next subscript on the grid K. This will be repeated until the critical value is exceeded.

We denote by $\hat{k}$ the first subscript $k \in K$ for which the critical value D is exceeded. Set $\widehat{s} = t_{\hat{k}}$, which is called in the sequel the *breaking point*. We aim to choose the adaptive threshold $\widehat{\tau}$ by maximizing the quasi-log-likelihood function

$$\begin{aligned} &\max_{\theta} \mathcal{L}^{part}(\theta \,|\, \mathbf{z}) - \mathrm{Pen}(\widehat{\theta}_{\hat{s}} \,\big|\, \mathbf{z}) \\ &\qquad = \mathcal{L}^{part}(\widehat{\theta}_{\tau} \,\big|\, \mathbf{z}) - \mathrm{Pen}(\widehat{\theta}_{\hat{s}} \,\big|\, \mathbf{z}), \end{aligned}$$

where $\mathrm{Pen}(\widehat{\theta}_{\hat{s}} \,\big|\, \mathbf{z})$ is the penalty function defined by

$$\mathrm{Pen}(\theta \,|\, \mathbf{z}) = \mathcal{L}^{part}(\theta \,|\, \mathbf{z}). \tag{3.4.3}$$

Taking into account (3.4.3), it follows that the second term of (3.4.2) can be viewed as the penalized quasi-log-likelihood

$$\mathcal{L}^{Pen}(s, \tau) = \mathcal{L}^{part}(\widehat{\theta}_{\tau} \,\big|\, \mathbf{z}) - \mathrm{Pen}(\widehat{\theta}_{s} \,\big|\, \mathbf{z}). \tag{3.4.4}$$

We find the subscript which maximize the penalized quasi-log-likelihood

$$\widehat{l} = \underset{\zeta' \hat{k} \le l \le (1-\zeta'')\hat{k}}{\mathrm{argmax}} \; \mathcal{L}^{Pen}(\hat{s}, t_l). \tag{3.4.5}$$

Finally, we set the adaptive threshold

$$\widehat{\tau} = t_{\hat{l}} \tag{3.4.6}$$

and its associated parameter $\widehat{\theta}_{\hat{\tau}}$.

3.5 Aggregation

The transition between the non-parametric part and the parametric part in the model (3.2.3) can sometimes be rough, especially when the sample size is small. We propose two ways of smoothing the transition relying on a aggregating estimators corresponding to different thresholds.

3.5.1 Simple aggregation

The first aggregation we describe can be called 'simple aggregation' as we aggregate the estimated cumulative hazard function from multiple thresholds. The procedure can be resumed with 3 simple steps, where the observations t_i are ordered in the decreasing order such that $t_1 > ... > t_n$.

- **Step 1.** Choose $M \geq 1$ thresholds $\tau_1 = t_{m_0}, \ldots, \tau_m = t_{m_0+M-1}$ from the observed values, where $m_0 \geq 1$.
- **Step 2.** For each chosen threshold τ_k compute the estimated cumulative hazard function $\widehat{H}_{z,\tau_k,\theta}(x)$.
- **Step 3.** Compute the simple aggregation estimator by

$$\widehat{S}_{\text{sa}}(\, x \,|\, z\,) = \exp\left(-\frac{1}{M}\sum_{k=1}^{M} \widehat{H}_{z,\tau_k,\theta}(x)\right). \tag{3.5.1}$$

For the algorithm to work, we need to choose m_0 as the first observation (censored or not) having at least one non-censored observation above it: $m_0 \geq \min\{m\colon \sum_{t_i>t_m} \delta_i = 1\}$. With $M = 1$ and $m_0 = k$, the procedure becomes the estimation of the semi-parametric model (3.2.3) with a fixed threshold $\tau = t_k$.

3.5.2 Adaptive aggregation

The second aggregation we describe can be called 'adaptive aggregation' as we aggregate cumulative hazard functions from the adaptive procedure described in Section 3.4. Let $l_1, \ldots, l_{N_{\hat{s}}}$ be the reversed ranks of the sequence

$$\mathcal{L}^{Pen}\,(t_{\hat{s}}, t_l)\,,\ \zeta'\hat{s} \leq l \leq (1-\zeta'')\hat{s},$$

where $N_{\hat{s}}$ is its cardinality and $\hat{s}$ is the breaking point computed by the adaptive procedure described in Section 3.4. For the adaptive aggregation we proceed in the same way as in the case of the simple aggregation described above. It can be resumed with the following steps:

- **Step 1.** Choose $M \geq 1$ thresholds $\tau_1 = t_{l_1}, \ldots, \tau_M = t_{l_M}$ from the observed values.
- **Step 2.** For each chosen threshold τ_{l_k}, compute the estimated cumulative hazard function $\widehat{H}_{z,\tau_{l_k},\theta}(x)$.
- **Step 3.** Compute the weights on the estimated cumulative hazard functions from the value of the penalized likelihood (3.4.4) by

$$w_{l_k} = \mathcal{L}^{Pen}\left(t_{\hat{s}}, t_{l_k}\right) / \sum_{i=1}^{M} \mathcal{L}^{Pen}\left(t_{\hat{s}}, t_{l_i}\right).$$

- **Step 4.** Compute the adaptive aggregation estimator by

$$\widehat{S}_{\text{aa}}(x \mid z) = \exp\left(-\sum_{k=1}^{M} w_{l_k} \widehat{H}_{z,\tau_{l_k},\theta}(x)\right). \tag{3.5.2}$$

Note that with $M = 1$, the procedure becomes the estimation of the semi-parametric model (3.2.3) with the adaptive threshold τ chosen from the procedure described in Section 3.4.

3.6 Bootstrap estimator

When the observations exceeding the threshold are few (this can be the case in heavy right censoring), the extremes observations have a great weight on the estimation of the Pareto parameter. We propose a bootstrap method in order to improve the estimation of the survival probabilities in the case of few extremes observations.

The basic idea is to randomly draw datasets of size n with replacement from the observed data. We repeat the process B times, producing B datasets. For each dataset, we calculate the cumulative hazard function of (3.2.8): $\widehat{H}^b_{\tau^b,\widehat{\theta}^b_\tau}(x \mid z)$, where $b = 1, ..., B$. Note that the threshold is computed with each bootstrap sample as well as the estimation of the Pareto parameter The bootstrap estimator of (3.2.8) is computed by :

$$\widehat{S}_{\text{boot}}(x \mid z) = \exp\left(-\sum_{b=1}^{B} \widehat{H}^b_{\tau^b,\widehat{\theta}^b_\tau}(x \mid z)\right). \tag{3.6.1}$$

The bootstrap method can also be used on the aggregation estimators described in Section 3.5.1 and 3.5.2. The bootstrap estimator of the "simple aggregation" is defined by:

$$\widehat{S}_{\text{boot,sa}}(x \mid z) = \exp\left(-\sum_{b=1}^{B} \widehat{H}^b_{\text{sa}}(x \mid z)\right). \tag{3.6.2}$$

The bootstrap estimator of the "adaptive aggregation" can be computed by:

$$\widehat{S}_{\text{boot,aa}}(x \mid z) = \exp\left(-\sum_{b=1}^{B} \widehat{H}^{b}_{\text{aa}}(x \mid z)\right). \tag{3.6.3}$$

The aggregation in the bootstrap method are done on the cumulative hazard function. In Section 3.7, we will compare this aggregation with the one performed on the survival function instead.

3.7 Simulation study

In this section, we present simulation studies to evaluate how our estimators perform. In a first part, we perform a simulation study on the choice of the threshold and compare it to other possible choice. In a second part, we compare all the estimators of the survival function against the usual Nelson-Aalen estimator with two different simulation distributions. We are interested in the values of the baseline survival probability $S_0(x)$ when x is large. To compare the estimations, we use the relative mean square error (RelMSE), which we define as $RelMSE_{\widehat{S}_0(x)} = \mathbb{E}\left(\ln^2 \frac{\widehat{S}_0(x)}{S_0(x)}\right)$. One can compare the estimated survival function and the true survival function for any z by multiplying the error for the baseline by $e^{2\beta.z}$. The parameters of the adaptive procedure in Section 3.4 are set to the following values $n_{\text{grid}} = 100$, $\zeta' = 0.25$, $\zeta'' = 0.05$. A simulation study has been performed in Durrieu et al. [54] on the choice of these parameters and led to these values.

Let be the transformed Cauchy distribution with location parameter x_0 and scale γ defined as:

$$F(x) = 1 - \frac{1 - \frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right) + \frac{1}{2}}{1 - \frac{1}{\pi}\arctan\left(\frac{0-x_0}{\gamma}\right) + \frac{1}{2}}$$

The critical value was chosen by simulation using the Pareto distribution which led us to set it to 10. Figure 3.2 shows the empirical distribution function of the test statistic on 10000 simulations of the Pareto distribution under the Cox model.

3.7.1 Choice of the threshold

The number of aggregation is set by simulations to $M_a = 50$ in the procedure described in Section 3.5.2.

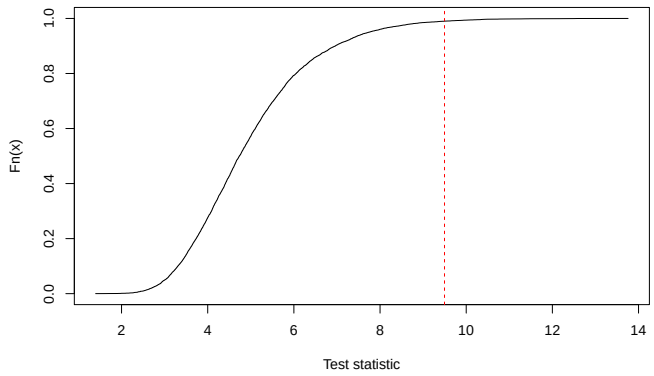

Figure 3.2: Empricial distribution function of the test statistic. The red dashed line shows the test statistic for the quantile of order 0.99.

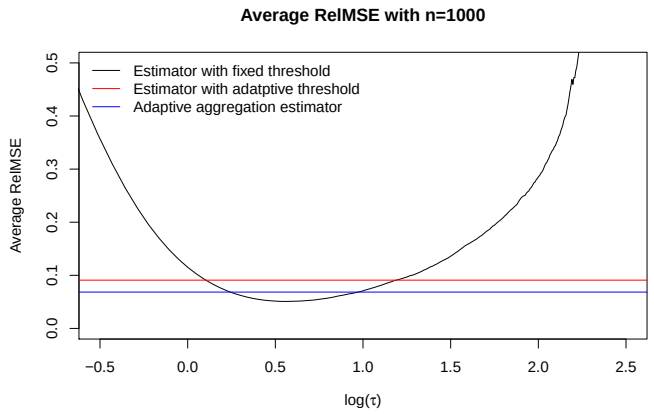

Figure 3.3: Average RelMSE's of the estimator with fixed threshold τ (black line), of the estimator with the adaptive choice of the threshold (red line) and of the adaptively aggregated estimator (blue line). All estimators are computed on 10000 simulations of the transformed Cauchy distribution with parameters $x_0 = 0$ and $\gamma = 1$ and with parameters $x_0 = 0$ and $\gamma = 2$ for the censorship .

The comparison with the "simple aggregation" estimator is not presented here to show how an adaptive selection of the threshold perform against a arbitrary choice. The performance of the adaptive choice of the threshold is shown on Figure 3.3.

We plot the average RelMSE of the estimated survival function (3.2.8) with fixed threshold τ, where τ is running on the uniform grid on $[0.1, 20]$ (black line) and the average RelMSE of the estimated survival function (3.2.8) with the adaptive

threshold $\widehat{\tau}$ (red line) chosen by the procedure described in 3.4. The average RelMSE is defined by : $ARelMSE_{\widehat{S}_0} = 1/n_{grid} \sum_{j=1}^{n_{grid}} RelMSE_{\widehat{S}_0(x_j)}$, where $x_j, j = 1, ..., n_{grid}$ is a geometric grid on $[0.1, 100]$. The average RelMSE of the aggregated estimated survival function $\widehat{S}_{0,\text{aa}}$ is also shown in blue line. One can see that the adaptive choice of the threshold is not optimal but it is close to the best one. Recall that the choice of the adaptive threshold is done without any prior knowledge of the 'best' choice of the threshold (which sometimes is also called 'oracle' choice).

3.7.2 Survival probabilities estimation

The number of aggregation M_a and M_s described in Section 3.5.1 and 3.5.2 has to be set beforehand. We chose a sample size of $n = 500$ in order to perform a Monte-Carlo study.

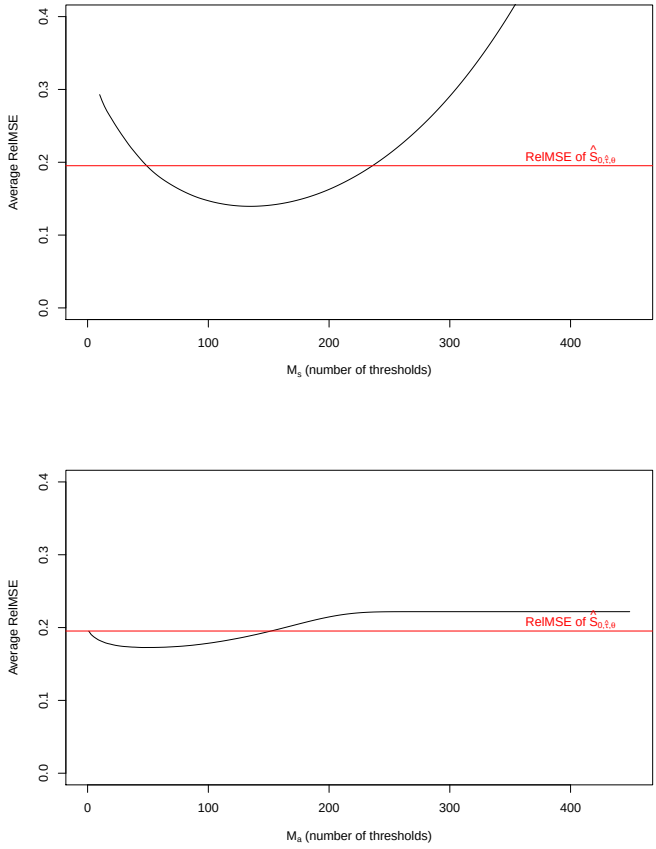

Figure 3.4: Average RelMSE for $\widehat{S}_{\text{sa},0}$ depending on the number of thresholds M_s (top). Average RelMSE for $\widehat{S}_{\text{aa},0}$ depending on the number of thresholds M_a (bottom). The RelMSE for $\widehat{S}_{0,\widehat{\tau},\theta}$ is added in red.

The number of threshold for which the minimum ARelMSE is attained is different for both of the methods. With the simple aggregation described in Section 3.5.1, the minimum is obtained for $M_s = 136$. Whereas it is obtained for $M_a = 47$ with the adaptive aggregation described in Section 3.5.2. The values are set in the rest of the study to $M_s = 135$ for the simple aggregation and $M_a = 50$ for the adaptive aggregation.

In order to illustrate the choice of the transformed Cauchy distribution with location $x_0 = 0$ and scale $\gamma = 1$, we show in Figure 3.5 the density of the failure time X, the density of the threshold $\widehat{\tau}$ and the breaking point $\widehat{s}$ chosen by the adaptive procedure presented in Section 3.4.

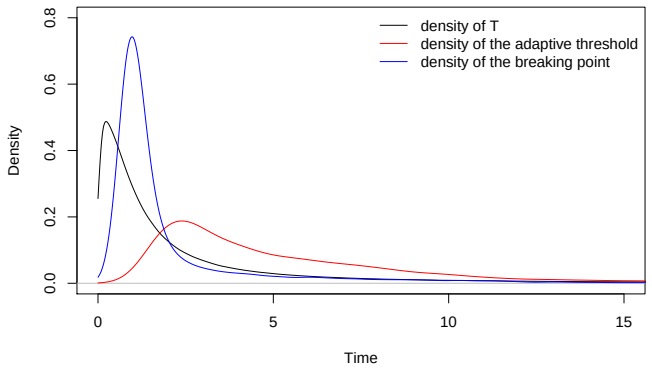

Figure 3.5: Density of the threshold $\widehat{\tau}$, the breaking point $\widehat{s}$ and the observations T computed on 10000 simulations of the transformed Cauchy distribution with parameters $x_0 = 0$ and $\gamma = 1$.

One can see that the breaking point is chosen after the location parameter of the transformed Cauchy distribution and the threshold is therefore set to a value greater than it.

Assume that we have decided on the sample size n, the parameter β, the covariate distribution of Z, the baseline survival function $S_0(\cdot)$ and the censoring distribution $S_C(\cdot \mid z)$. We generated data sets in our simulation study following the pattern : we consider F_0 to follow the transformed Cauchy distribution with parameters $x_0 = 0$ and $\gamma = 1$. In the first study, the survival distribution function S_C is assumed following the Cauchy distribution with parameter $x_0 = 0$ and $\gamma = 2$. In a second study, S_C is assumed following the transformed Cauchy distribution with parameter $x_0 = 10$ and $\gamma = 0.1$. In both cases, the censoring survival distribution function doesn't depend on the covariates. The mean censoring rate is around 50% for the first distribution of S_C and around 20% for the second distribution of S_C.

$n = 100$					
x	100	200	300	400	500
RelMSE of $\widehat{S}_0(x)$	4.6750	7.8920	10.2147	12.0651	13.6145
RelMSE of $\widehat{S}_{0,\widehat{\tau},\beta}(x)$	1.5908	2.1641	2.5413	2.8277	3.0605
RelMSE with simple aggregation	1.0306	1.4164	1.6714	1.8654	2.0234
RelMSE with adaptive aggregation	1.2361	1.6976	2.0023	2.2341	2.4229
$n = 500$					
x	100	200	300	400	500
RelMSE of $\widehat{S}_0(x)$	2.0733	4.1024	5.7537	7.1133	8.2953
RelMSE of $\widehat{S}_{0,\widehat{\tau},\beta}(x)$	0.4209	0.5844	0.6928	0.7754	0.8427
RelMSE of $\widehat{S}_{sa,0}(x)$	0.4149	0.5771	0.6846	0.7666	0.8335
RelMSE of $\widehat{S}_{aa,0}(x)$	0.3763	0.5253	0.6244	0.6999	0.7616
RelMSE of $\widehat{S}_{\text{boot},0,\widehat{\tau},\beta}(x)$	0.7998	1.1391	1.3659	1.5396	1.6816
RelMSE of $\widehat{S}_{\text{boot},sa,0}(x)$	0.3754	0.5242	0.6230	0.6985	0.7601
RelMSE of $\widehat{S}_{\text{boot},aa,0}(x)$	0.6910	0.9782	1.1695	1.3157	1.4351

Table 3.1: 1000 Monte-Carlo simulations where the parameters of the censorship distribution are $x_0 = 0$ and $\gamma = 2$.

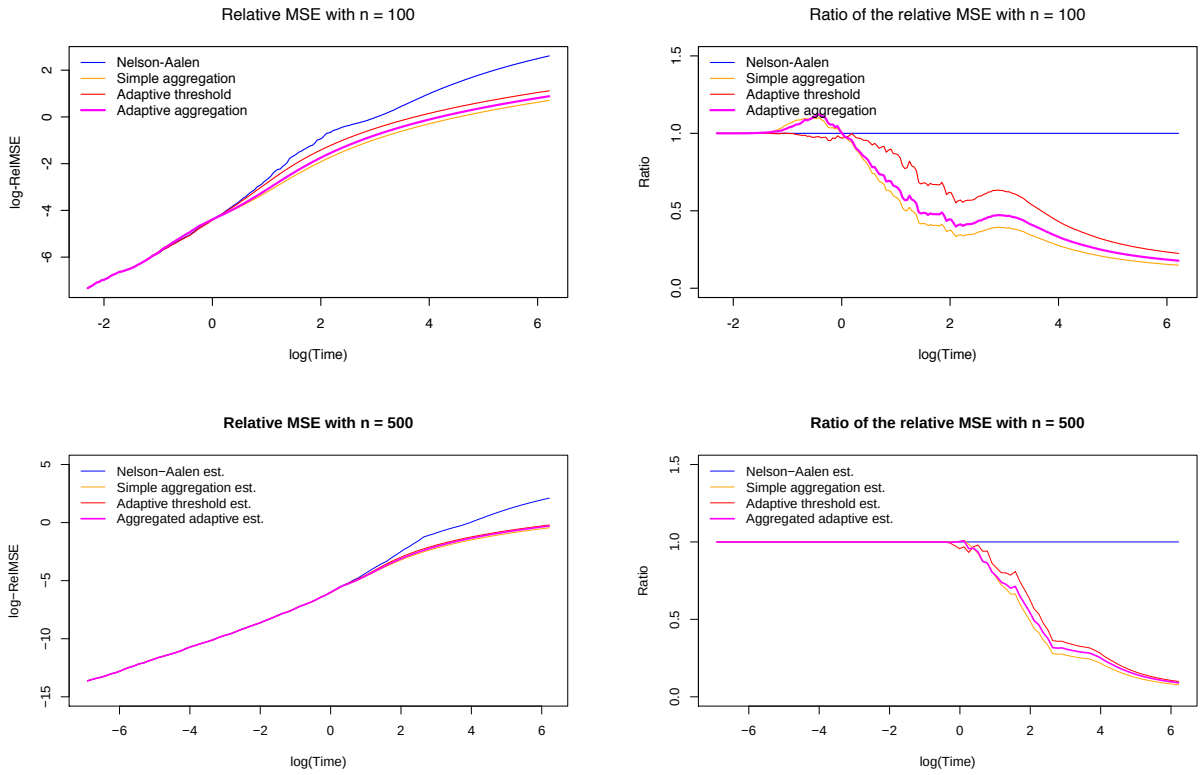

Figure 3.6: Relative MSE (left) and ratio of the relative MSE (right) for $n = 100$ (top) and $n = 500$ (bottom) and $NMC = 1000$. The parameters of the censorship distribution are $x_0 = 0$ and $\gamma = 2$.

The covariate is supposed to be a random uniform variable in $[-1, 1]$ and the parameter β is set to -0.5. Table 3.1 gives the results of a Monte-Carlo simulation for the estimation of the baseline function with S_C following a transformed Cauchy distribution with parameters $x_0 = 0$ and $\gamma = 2$, this information is plotted in Figure 3.6.Table 3.2 gives the results when S_C follows a transformed Cauchy distribution with parameters $x_0 = 10$ and $\gamma = 0.1$.

$n = 100$					
x	100	200	300	400	500
RelMSE of $\widehat{S}_0(x)$	5.1196	8.6637	11.1824	13.1688	14.8236
RelMSE of $\widehat{S}_{0,\widehat{\tau},\beta}(x)$	0.9857	1.3812	1.6447	1.8460	2.0104
RelMSE with simple aggregation	0.7266	1.0534	1.2749	1.4456	1.5857
RelMSE with adaptive aggregation	0.5970	0.8203	0.9681	1.0807	1.1725
$n = 500$					
x	100	200	300	400	500
RelMSE of $\widehat{S}_0(x)$	4.4576	7.7855	10.1777	12.0743	13.6595
RelMSE of $\widehat{S}_{0,\widehat{\tau},\beta}(x)$	0.2162	0.3128	0.3780	0.4281	0.4691
RelMSE with simple aggregation	0.4499	0.7033	0.8787	1.0154	1.1283
RelMSE with adaptive aggregation	0.1597	0.2257	0.2698	0.3036	0.3313
RelMSE of $\widehat{S}_{\text{boot},0,\widehat{\tau},\beta}(x)$	1.3577	2.1829	2.7577	3.2067	3.5786
RelMSE of $\widehat{S}_{\text{boot},sa,0}(x)$	0.2822	0.4292	0.5303	0.6089	0.6737
RelMSE of $\widehat{S}_{\text{boot},aa,0}(x)$	0.1976	0.2871	0.3476	0.3941	0.4322

Table 3.2: 1000 Monte-Carlo simulations with the following parameters of the censorship distribution $x_0 = 10$ and $\gamma = 0.1$.

The RelMSE of the bootstrap estimate is shown by curiosity in Table 3.1 and Table 3.2 for $n = 500$, we can see that it does not perform poorly. The bootstrap estimate can't be calculated on $n = 100$ because of the censoring rate.

We have added another model based on the log-Gamma law, which has a rather "bad" slowly varying part of type $\log x$. The results are reported in Figure 3.7 and Table 3.3. These results show that the introduction of the aggregated estimator improves the estimation of the tails significantly with respect to standard Nelson-Aalen estimator. In particular, for low sample sizes, the aggregation improves the error estimation over the simple adaptive choice. The simulated distribution of the survival and censoring time follow a log-gamma distribution where the density function is defined for $x > 1$ by:

$$f(x) = \frac{b^a \ln(x)^{a-1}}{\Gamma(a) x^{b+1}},$$

where $a > 0$ is the shape parameter, $b > 0$ is the rate parameter and $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$ is the Gamma function.

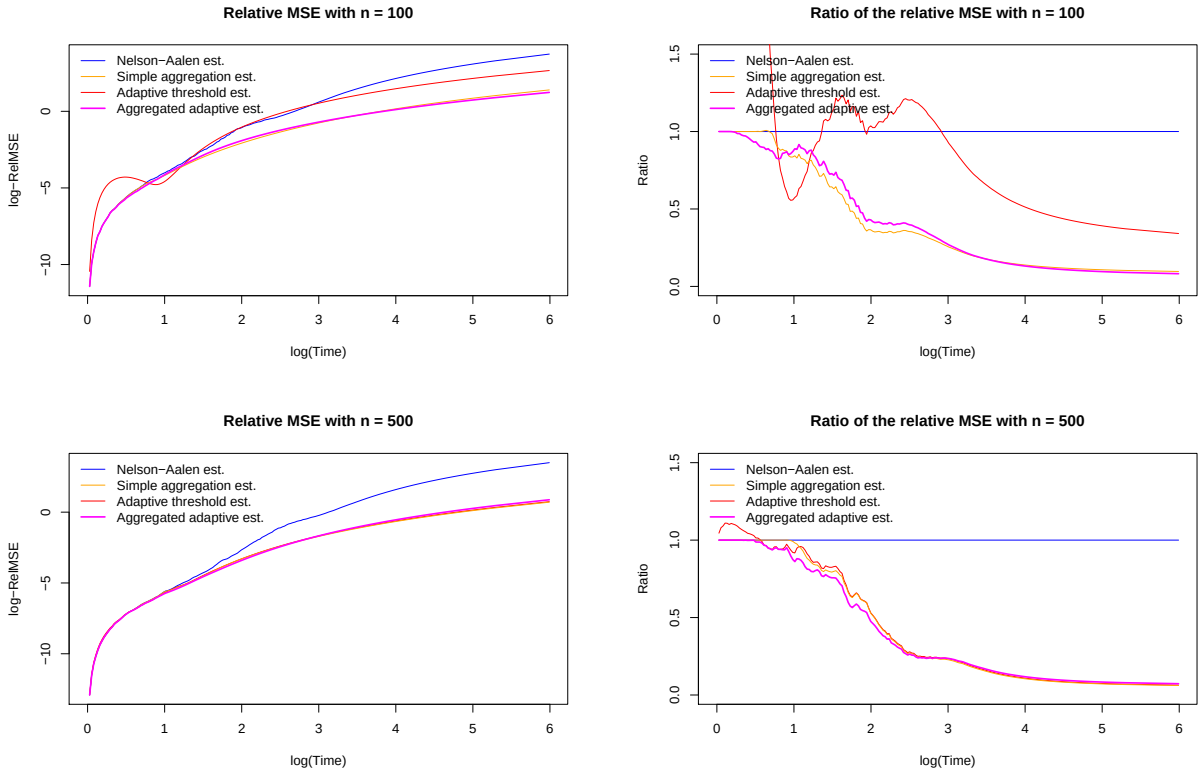

Figure 3.7: Relative MSE (left) and ratio of the relative MSE (right) for $n = 100$ (top) and $n = 500$ (bottom) and $NMC = 1000$. The censorship follows a log-Gamma distribution with $a = 5$ and $b = 3.5$.

$n = 100$					
x	100	200	300	400	500
RelMSE of $\widehat{S}_0(x)$	3.7469	7.6762	15.7922	27.1767	41.9159
RelMSE of $\widehat{S}_{0,\widehat{\tau},\beta}(x)$	2.5996	4.0775	6.7142	10.1159	14.3147
RelMSE with simple aggregation	0.7026	1.0984	1.8291	2.8042	4.0424
RelMSE with adaptive aggregation	0.7087	1.0536	1.6645	2.4551	3.4391
$n = 500$					
x	100	200	300	400	500
RelMSE of $\widehat{S}_0(x)$	1.7343	4.3083	10.5811	20.0979	32.9481
RelMSE of $\widehat{S}_{0,\widehat{\tau},\beta}(x)$	0.2985	0.4961	0.8774	1.4057	2.0957
RelMSE with simple aggregation	0.2876	0.4771	0.8439	1.3538	2.0217
RelMSE with adaptive aggregation	0.3110	0.5347	0.9744	1.5906	2.4005

Table 3.3: 1000 Monte-Carlo simulations with the censorship distribution following a log-Gamma distribution with $a = 5$ and $b = 3.5$.

We considered F_0 to follow a log-Gamma distribution with parameters $a = 2$ and $b = 2$ and S_C to follow a log-Gamma distribution with parameters $a = 5$ and $b = 3.5$. The mean censoring rate is around 60%. Figure 3.7 shows the $RelMSE$ and the ratio of $RelMSE$ of the baseline function compared to the Nelson-Aalen estimator where the baseline function follows a log-Gamma distribution with parameters $a = 2$ and $b = 2$. In Table 3.3, we compared the $RelMSE$ of estimated survival probabilites for extremes values.

3.8 Applications

3.8.1 Bladder data set

As a second example, we consider the data set `bladder` included in the R package `survival` (https://cran.r-project.org/web/packages/survival/index.html). This data set concerns the comparison of different treatments on the recurrence of Stage I bladder tumor (see [94] for more details). We study here only the difference between the placebo and the thiotepa treatment. The initial purpose behind the study of this data set was to determine if the treatment had an effect on the recurrence of the bladder tumor. This study has been done in [95] using the usual Cox model. We want to extend the problem by determining the probability of having the first recurrence of the bladder tumor (the first recurrence is the most important to examine the treatment effect) at the end of the study for the placebo and treatment groups or at what time does the estimated probability of having the first recurrence fall below 0.3.

We consider the observed time as the time between two recurrences or between the last recurrence and the censoring time. The covariate includes the treatment, the number of initial tumors, the size of the initial tumor and the number of recurrences.

One can see on the Figure 3.8 that our model fits the tail of the distribution and the recurrence time is estimated after the last observed time. The estimated survival probability of having the first recurrence of the tumor beyond 3 and 4 years are given in the following table.

The next table gives the estimated time at which a patient has a survival probability of having a first recurrence of 0.3 and 0.4.

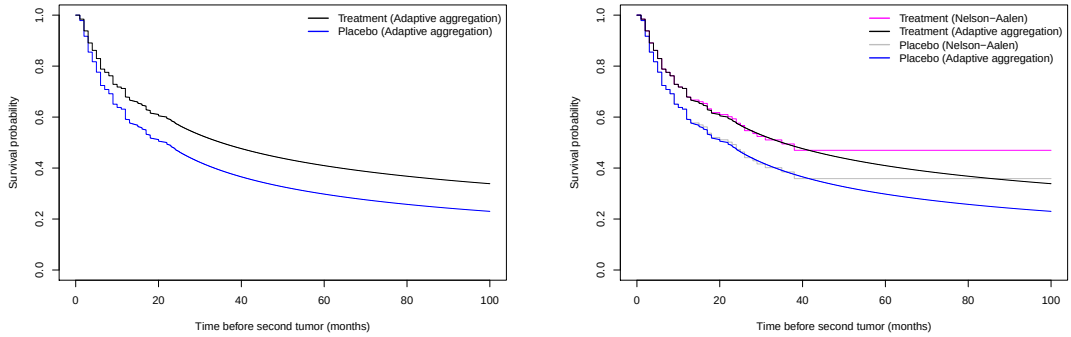

Figure 3.8: Estimated survival probabilities of the time of having the first recurrence of the bladder tumor when the initial size of the tumor is 1 and the initial number of tumors is 1.

Time (months)	72	96
Nelson-Aalen estimator (Placebo)	0.3586	0.3586
Nelson-Aalen estimator (Treatment)	0.4697	0.4697
Adaptive aggregation estimator (Placebo)	0.2715	0.2348
Adaptive aggregation estimator (Treatment)	0.3826	0.3438

Table 3.4: Estimated probabilities of having the first recurrence of the tumor beyond 3 and 4 years when the initial size of the tumor is 1, the initial number of tumors is 1.

Survival probability	0.3	0.4
Nelson-Aalen estimator (Placebo)	NA	35
Nelson-Aalen estimator (Treatment)	NA	NA
Adaptive aggregation estimator (Placebo)	95.6232	43.3311
Adaptive aggregation estimator (Treatment)	194.9080	66.58699

Table 3.5: Estimated time (months) of having the first recurrence of the tumor with a survival probabilities of 0.3 and 0.4 when the initial size of the tumor is 1, the initial number of tumors is 1.

One can see that with the model proposed in this paper, the initial problem of analyzing the estimated regression parameter to observe an effect of the treatment has been extended. It is possible to give an estimated probability of having the recurrence before a certain time and it's possible to give an estimated time of recurrence for a given probability.

3.8.2 Application to electric consumption prediction

In order to offer an alternative to load shedding in case of an electric constraint, a research project has been conducted in Lorient, France, to study the electric consumption of households. The idea is to allocate the available electricity among all the consumers in the concerned area by lowering their usual power and so, avoiding the black out. This action can be done remotely from the smart meter. One of the objectives of the experiment is to study the behaviour of the consumers and the consequences on their consumption. Of course, the process will be used only when the market offers can not cope anymore. This action is known as 'active power modulation'. The data are collected on selected houses to study the effect during the electric constraint. For example, if a house with a maximal electric power contract of 9 kiloVolt Ampere has a constraint of 50%, the maximal electric power becomes 4.5 kVA. The goal of this study is to minimise the number of houses without electricity during a major power outages. If the electric power requested by the house exceeds the maximal permitted power, the breaker cuts off and the house has no electricity. In this section, we predict the electric power level for one random house during the time of the constraint and compare the measured level with what really happened.

The data used in this application are the electric power of a house with a maximal power contract of 9 kVA. A measurement of the electric power is made every 10 minutes and corresponds to the mean load power requested in 10 minutes. The outside temperature is collected at the same times. The study period started on the 23rd December, 2015 and finished on the 21st March, 2016. Figure 3.9 shows the consumption of the studied house during the period and the measured outside temperature.

As one can see in the Figure 3.9, we deal with a time series. We decided to remove the dependence of the time by discretized the data by hour, under the hypothesis that during the winter, the distribution of the electric power during the same hours remains similar over the days. The hours become part of the covariate and a binary information is given, e.g. a measurement between 2h et 3h will have a 1 in during this hour and 0 elsewhere.

Moreover, the temperature is included in the covariate with a subtle transformation. Indeed, we separated the temperature into 4 linear covariates as we assume that the parameter of the temperature will not be constant over the scope of the temperature.

For this data set, the cumulative hazard rate functions are not proportionals. We decided to separate the data into five classes to improve the estimation of extreme probabilities. For each group corresponds a period during the day. The hour classes are from 22h until 6h which corresponds to the night. From 6h until 10h, which corresponds to the morning. From 10h until 14h, which corresponds to the lunch time. From 14h until 18h, which corresponds to the afternoon and finally from 18h

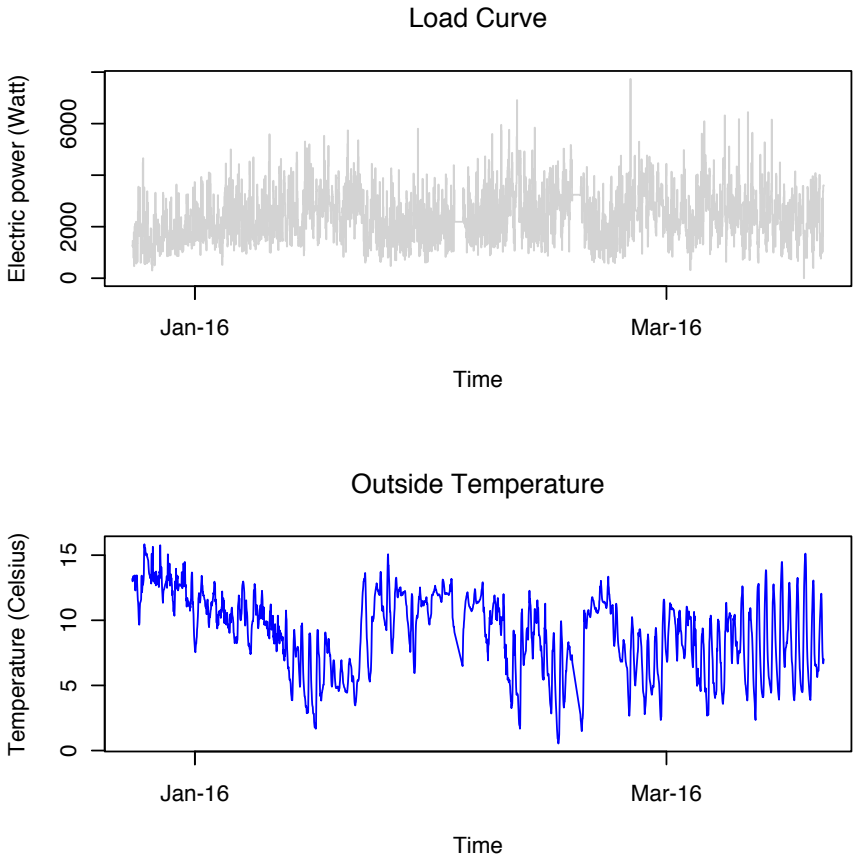

Figure 3.9: Load curve (top) and outside temperature (bottom) for the studied period.

until 22h, which corresponds to the evening.

The proportional hazards assumption almost holds for each groups. We are interested in the impact of the temperature onto this assumption, but the size of the data is not big enough to verify this.

Using the hypothesis from which the distribution of the electric power during the same hours is similar over the days during the winter, we estimate the survival functions. The goal is to predict the probability to exceed the maximal authorized power during the time of the constraint. Table 3.6 shows the different time of the constraint, the value of the maximal power during the constraint and the average outside temperature during the period starting the 23rd December, 2015 and finishing the 21st March, 2016. Recall that the maximal power of this house when there is no constraint is 9 kVA. For each constraint, we give the estimated survival probability to exceed the maximal power given the time and the outside temperature.

Constraint day **Constraint hours**	11th January 17-19h	13th January 14-18h	18th January 14-18h	25th February 14-18h
Maximal power	6.3 kVA	4.5 kVA	4.5 kVA	4.5 kVA
Outside temperature	8.56 TÂ°C	8.85 TÂ°C	6.82 TÂ°C	9.85 TÂ°C
Estimated survival probability	0.0015	0.09	0.1188	0.0778
Number of cut off of the breaker	0	0	0	0
Constraint day **Constraint hours**	1st March 17-19h	7th March 14-18h	18th March 10-12h	
Maximal power	3.6 kVA	2.8 kVA	2.8 kVA	
Outside temperature	10.92 TÂ°C	9.70 TÂ°C	9.98 TÂ°C	
Estimated survival probability	0.0526	0.5018	0.7101	
Number of cut off of the breaker	1	2	8	

Table 3.6: Date of the electric constraint, maximal power, outside temperature and estimated survival probabilities during the electric constraint. The number of cut off of the breaker represents the number of time the breaker cut off during the constraint period.

The probability corresponding to a return period of 4 hours (happen once in any given 4 hours period) is $\frac{1}{24} \simeq 0.04$. For a return period of 2 hours, the corresponding probability is 0.08. We can see on Table 3.6 that we detect five estimated probabilities exceeding the probability associated to the return period. These probabilities correspond to the constraints of the 13th of January, the 18th of January, the 15th of February, the 7th of March and the 18th of March. This house is therefore considered at risk during these five constraints.

We are now interested to see what really happened for this house during these constraints. This house had one opening of its breaker during the constraint of the 1st of March, two during the constraint of the 7th of March, eight during the constraint of the 18th of March and none during the other constraints. We can see that we have very high estimated probabilities of having one observation exceeding the maximal power during constraint for the 7th and 18th of March and that the house had multiple cuts off during these constraints. We can suppose that during the other constraints, the house anticipated the constraint period and reduced its electric power by changing its behavior.

3.9 Conclusion

In this chapter, we propose an extension of the Cox model in order to estimate probabilities of rare events and extreme quantiles. The model is semi-parametric and composed of the Nelson-Aalen estimator for the non-parametric part and the parametric part is described by a Pareto distribution. We prove the consistency of the estimator of the Pareto parameter and give an explicit convergence rate for the Hall model.

A data-driven choice of the threshold motivated by a goodness-of-fit test is proposed. An aggregated estimator and an adaptive aggregated estimator are suggested and studied in order to improve the fitness of the model onto the data. The performance of the proposed estimators is demonstrated on artificial data.

Two applications on real data sets are given. The application on the bladder data shows the motive of the model as it allows an estimation of extreme quantiles which was not possible with the usual Cox model. The application on the electric consumption gives an application onto data where the main purpose is to estimate survival probabilities and we are not interested to test if there is an effect of a treatment.

3.A Proofs of the results.

3.B Proofs of the results.

3.B.1 Proof of Theorem 3.2.1.

In the sequel we denote quasi-log-likelihood ratio by

$$\mathcal{L}(\theta', \theta \,|\, \mathbf{z}) := \mathcal{L}(\theta' \,|\, \mathbf{z}) - \mathcal{L}(\theta \,|\, \mathbf{z}) = \sum_{i=1}^{n} \ln \frac{dP_{S_{0,\tau,\theta'}}}{dP_{S_{0,\tau,\theta}}} (t_i, \delta_i \,|\, z_i) \,.$$

Recall that we denoted by $\mathbb{P}$ the probability measure corresponding to the "true" model which has the baseline survival function S_0.

Lemma 3.B.1. *For any $\theta, \theta' \in \mathbb{R}$, $\mathbf{z} \in \mathbb{Z}$ and any $x > 0$, it holds*

$$\mathbb{P}(\mathcal{L}(\theta', \theta \,|\, \mathbf{z}) > x + \sum_{i=1}^{n} \chi^2(P_{S_0}(. \,|\, z_i), P_{S_{0,\tau,\theta}}(. \,|\, z_i))) \leq e^{-x/2}$$

Proof. Let $P_{S_{0,\tau,\theta'}}(. \,|\, z_i)$ and $P_{S_{0,\tau,\theta}}(. \,|\, z_i)$ be the conditional cumulative distribution function of (T, Δ) given $Z = z_i$ where the survival function has a Pareto tail with

parameter θ' and θ respectively for a threshold τ. By Chebychev's exponential inequality, we have for any $y > 0$:

$$\begin{aligned}\mathbb{P}\left(\mathcal{L}(\theta', \theta \,|\, \mathbf{z}) > y\right) &\leq e^{-y/2}\mathbb{E}\left(e^{\frac{1}{2}\mathcal{L}(\theta', \theta \,|\, \mathbf{z})}\right) \\ &\leq \exp\left[-y/2 + \ln\left(\mathbb{E}\left(e^{\frac{1}{2}\mathcal{L}(\theta', \theta \,|\, \mathbf{z})}\right)\right)\right].\end{aligned}$$

As the triplet $\{t_1, \delta_1 \,|\, z_1\}, ..., \{t_n, \delta_n \,|\, z_n\}$ are independent, we can write the term $\ln\left(\mathbb{E}\left(e^{\frac{1}{2}\mathcal{L}(\theta', \theta \,|\, \mathbf{z})}\right)\right)$ as

$$\begin{aligned}\ln\left(\mathbb{E}\left(e^{\frac{1}{2}\mathcal{L}(\theta', \theta \,|\, \mathbf{z})}\right)\right) &= \ln\left(\mathbb{E}\left(e^{\frac{1}{2}\sum_{i=1}^{n} \ln \frac{dP_{S_{0,\tau,\theta'}}}{dP_{S_{0,\tau,\theta}}}(t_i, \delta_i \,|\, z_i)}\right)\right) \\ &= \ln\left(\mathbb{E}\left(\prod_{i=1}^{n} \sqrt{\frac{dP_{S_{0,\tau,\theta'}}}{dP_{S_{0,\tau,\theta}}}(t_i, \delta_i \,|\, z_i)}\right)\right) \\ &= \sum_{i=1}^{n} \ln\left(\mathbb{E}\left(\sqrt{\frac{dP_{S_{0,\tau,\theta'}}}{dP_{S_{0,\tau,\theta}}}(t_i, \delta_i \,|\, z_i)}\right)\right).\end{aligned}$$

By Hölder's inequality, we have

$$\begin{aligned}\mathbb{E}\left(\sqrt{\frac{dP_{S_{0,\tau,\theta'}}}{dP_{S_{0,\tau,\theta}}}(t_i, \delta_i \,|\, z_i)}\right) &\leq \sqrt{\mathbb{E}\left(\frac{dP_{S_{0,\tau,\theta'}}}{dP_{S_{0,\tau,\theta}}}(t_i, \delta_i \,|\, z_i)\right)} \\ &\leq \sqrt{1 + \chi^2(P_{S_0}(\cdot \,|\, z_i), P_{S_{0,\tau,\theta}}(\cdot \,|\, z_i))},\end{aligned}$$

where the χ^2 entropy between two equivalent probability measure is defined by (3.2.11). Then,

$$\begin{aligned}\mathbb{P}\left(\mathcal{L}(\theta', \theta \,|\, \mathbf{z}) > y\right) &\leq \exp\left[-\frac{y}{2} + \frac{1}{2}\sum_{i=1}^{n} \ln\left(1 + \chi^2(P_{S_0}(\cdot \,|\, z_i), P_{S_{0,\tau,\theta}}(\cdot \,|\, z_i))\right)\right] \\ &\leq \exp\left[-\frac{y}{2} + \frac{1}{2}\sum_{i=1}^{n} \chi^2(P_{S_0}(\cdot \,|\, z_i), P_{S_{0,\tau,\theta}}(\cdot \,|\, z_i))\right].\end{aligned}$$

Setting $y = x + \sum_{i=1}^{n} \chi^2(P_{S_0}(\cdot \,|\, z_i), P_{S_{0,\tau,\theta}}(\cdot \,|\, z_i))$, gives Lemma 3.B.1. □

Recall that the Kullback-Leibler divergence $\mathcal{K}(\theta', \theta)$ between two Pareto distribution with parameters θ' and θ is defined by (3.2.10). The following lemma gives the rate of convergence for $\widehat{\theta}_\tau$. We adapt the proof from [89] to the case of the Cox model under consideration in this paper.

Lemma 3.B.2. *For any $\theta > 0$, $\tau > x_0$ and $v > 0$, it holds*

$$\mathbb{P}\left(\widehat{n}_\tau \mathcal{K}(\widehat{\theta}_\tau, \theta) > v + \sum_{i=1}^{n} \chi^2(P_{S_0}(\cdot \,|\, z_i), P_{S_{0,\tau,\theta}}(\cdot \,|\, z_i)) + 2\ln(n)\right) \leq 2\exp(-v/2),$$

where $\widehat{n}_\tau = \sum_{t_i > \tau} \delta_i$.

Proof. It is easy to see that the assertion of lemma follows from the following bound:

$$\mathbb{P}\left(\widehat{n}_\tau \mathcal{K}(\widehat{\theta}_\tau, \theta) > v + \sum_{i=1}^{n} \chi^2(P_{S_0}(\cdot \mid z_i), P_{S_{0,\tau,\theta}}(\cdot \mid z_i))\right) \leq 2n \exp\left(-\frac{v}{2}\right). \quad (3.B.1)$$

The likelihood ratio is given by

$$\begin{aligned}\mathcal{L}(\theta', \theta \mid \mathbf{z}) &:= \mathcal{L}(\theta' \mid \mathbf{z}) - \mathcal{L}(\theta \mid \mathbf{z}) \\ &= \sum_{i=1}^{n} \ln p_{S_{0,\tau,\theta'}}(t_i, \delta_i \mid z_i) - \sum_{i=1}^{n} \ln p_{S_{0,\tau,\theta}}(t_i, \delta_i \mid z_i).\end{aligned} \quad (3.B.2)$$

Then, removing the censoring part from the likelihood ratio, we have

$$\begin{aligned}\mathcal{L}(\theta', \theta \mid \mathbf{z}) &= \sum_{i=1}^{n} \ln(h_{z_i,\tau,\theta'}(t_i)^{\delta_i} S_{z_i,\tau,\theta'}(t_i)) - \sum_{i=1}^{n} \ln(h_{z_i,\tau,\theta}(t_i)^{\delta_i} S_{z,\tau,\theta}(t_i)) \\ &= \sum_{i=1}^{n} \delta_i \ln(h_{0,\tau,\theta'}(t_i) e^{\beta \cdot z_i}) + \ln(S_{z_i,\tau,\theta'}(t_i)) \\ &\quad - \delta_i \ln(h_{0,\tau,\theta}(t_i) e^{\beta \cdot z_i}) - \ln(S_{z_i,\tau,\theta}(t_i \mid z_i))).\end{aligned}$$

Developing the terms, we obtain

$$\begin{aligned}\mathcal{L}(\theta', \theta \mid \mathbf{z}) = \sum_{i=1}^{n} &\delta_i \ln(h_0(t_i)) \mathbb{1}_{t_i \leq \tau} + \delta_i \ln(\frac{1}{\theta' t_i}) \mathbb{1}_{t_i > \tau} + \delta_i \beta \cdot z_i \\ &+ e^{\beta \cdot z_i} \ln(S_{0,\tau,\theta'}(t_i)) \\ &- \delta_i \ln(h_0(t_i)) \mathbb{1}_{t_i \leq \tau} - \delta_i \ln(\frac{1}{\theta t_i}) \mathbb{1}_{t_i > \tau} - \delta_i \beta \cdot z_i - e^{\beta \cdot z_i} \ln(S_{0,\tau,\theta}(t_i)),\end{aligned}$$

and further

$$\begin{aligned}\mathcal{L}(\theta', \theta \mid \mathbf{z}) &= \sum_{i=1}^{n} \left[\delta_i \ln\left(\frac{\theta}{\theta'}\right) - e^{\beta \cdot z_i} \frac{1}{\theta'} \ln\left(\frac{t_i}{\tau}\right) + e^{\beta \cdot z_i} \frac{1}{\theta} \ln\left(\frac{t_i}{\tau}\right)\right] \mathbb{1}_{t_i > \tau} \\ &= \sum_{i=1}^{n} \left[\delta_i \ln\left(\frac{\theta}{\theta'}\right) + e^{\beta \cdot z_i} \ln\left(\frac{t_i}{\tau}\right)\left(\frac{1}{\theta} - \frac{1}{\theta'}\right)\right] \mathbb{1}_{t_i > \tau}.\end{aligned}$$

Since

$$\widehat{\theta}_\tau = \frac{\sum_{i:t_i > \tau} e^{\beta \cdot z_i \ln\left(\frac{t_i}{\tau}\right)}}{\sum_{i:t_i > \tau} \delta_i},$$

using (3.B.2), it follows that

$$\begin{aligned}\mathcal{L}(\theta', \theta \mid \mathbf{z}) &:= \mathcal{L}(\theta' \mid \mathbf{z}) - \mathcal{L}(\theta \mid \mathbf{z}) \\ &= \sum_{i=1}^{n} \mathbb{1}_{t_i > \tau} \delta_i \ln\left(\frac{\theta}{\theta'}\right) + \sum_{i=1}^{n} \mathbb{1}_{t_i > \tau} \delta_i \widehat{\theta}_\tau \left(\frac{1}{\theta} - \frac{1}{\theta'}\right) \\ &= \sum_{i=1}^{n} \mathbb{1}_{t_i > \tau} \delta_i \left[\ln\left(\frac{\theta}{\theta'}\right) + \widehat{\theta}_\tau \left(\frac{1}{\theta} - \frac{1}{\theta'}\right)\right] \\ &= \widehat{n}_\tau \left(\ln\left(\frac{\theta}{\theta'}\right) + \left(\frac{1}{\theta} - \frac{1}{\theta'}\right) \widehat{\theta}_\tau\right) \\ &= \widehat{n}_\tau \Lambda(\theta'),\end{aligned} \quad (3.B.3)$$

where we denoted for brevity $\Lambda(\theta') = \ln\left(\frac{\theta}{\theta'}\right) - \left(\frac{1}{\theta'} - \frac{1}{\theta}\right)\widehat{\theta}_\tau$. Since $\mathcal{K}(\theta', \theta) = \frac{\theta'}{\theta} - 1 - \ln(\frac{\theta'}{\theta})$, we then have the identity

$$\mathcal{K}(\widehat{\theta}_\tau, \theta) = \Lambda(\widehat{\theta}_\tau). \tag{3.B.4}$$

Using (3.B.3) and (3.B.4) we hope to bound $\mathcal{K}(\widehat{\theta}_\tau, \theta)$ in probability by Lemma 3.B.1. The problem is that $\widehat{\theta}_\tau$ is random and therefore we cannot apply Lemma 3.B.1 directly. We shall circumvent this difficulty in the following way. For any $y > 0$ and any $k \geq 1$, the inequality $k\mathcal{K}(u, \theta) > y$ can be equivalently written as

$$\left(\frac{1}{u} - \frac{1}{\theta}\right)\widehat{\theta}_\tau < -\frac{y}{k} + \ln\left(\frac{\theta}{u}\right).$$

Setting $g(u, k) = \frac{\ln\left(\frac{\theta}{u}\right) - \frac{y}{k}}{\left(\frac{1}{u} - \frac{1}{\theta}\right)}$, we have $g(u, k) > \widehat{\theta}_\tau$, when $0 < u < \theta$ and $g(u, k) < \widehat{\theta}_\tau$, when $u > \theta$. Moreover, the function $g(u, k)$ has a maximum for $0 < u < \theta$ and a minimum for $u > \theta$. When $0 < u < \theta$, we have $\lim\limits_{u \to 0^+} \frac{dg(u,k)}{du} = +\infty$ and $\lim\limits_{u \to \theta^-} \frac{dg(u,k)}{du} = -\infty$. When $u > \theta$, we have $\lim\limits_{u \to \theta^+} \frac{dg(u,k)}{du} = +\infty$ and $\lim\limits_{u \to +\infty} \frac{dg(u,k)}{du} = 0^-$. Let

$$\theta^+(k) = \operatorname*{argmax}_{0<u<\theta} g(u, k) \quad \text{and} \quad \theta^-(k) = \operatorname*{argmin}_{u>\theta} g(u, k),$$

then

$$\begin{aligned}
\{\widehat{n}_\tau \Lambda(\widehat{\theta}_\tau) > y, \widehat{\theta}_\tau < \theta\} &= \{g(\widehat{\theta}_\tau, \widehat{n}_\tau) > \widehat{\theta}_\tau, \widehat{\theta}_\tau < \theta\} \\
&\subset \{g(\theta^+(\widehat{n}_\tau), \widehat{n}_\tau) > \widehat{\theta}_\tau, \widehat{\theta}_\tau < \theta\} \\
&= \{\widehat{n}_\tau \Lambda(\theta^+(\widehat{n}_\tau)) > y, \widehat{\theta}_\tau < \theta\} \\
&\subset \{\widehat{n}_\tau \Lambda(\theta^+(\widehat{n}_\tau)) > y\}.
\end{aligned}$$

In the same way, we have $\{\widehat{n}_\tau \Lambda(\widehat{\theta}_\tau) > y, \widehat{\theta}_\tau > \theta\} \subset \{\widehat{n}_\tau \Lambda(\theta^-(\widehat{n}_\tau)) > y\}$. Since $\mathcal{K}(\widehat{\theta}_\tau, \theta) = \Lambda(\widehat{\theta}_\tau)$, this implies

$$\{\widehat{n}_\tau \Lambda(\widehat{\theta}_\tau) > y\} \subset \{\widehat{n}_\tau \Lambda(\theta^+(\widehat{n}_\tau)) > y\} \cup \{\widehat{n}_\tau \Lambda(\theta^-(\widehat{n}_\tau)) > y\}.$$

We then have

$$\begin{aligned}
\mathbb{P}(\widehat{n}_\tau \mathcal{K}(\widehat{\theta}_\tau, \theta) > y) &\leq \mathbb{P}(\widehat{n}_\tau \Lambda(\theta^+(\widehat{n}_\tau)) > y) + \mathbb{P}(\widehat{n}_\tau \Lambda(\theta^-(\widehat{n}_\tau)) > y) \\
&\leq \sum_{k=1}^{n} \mathbb{P}(\widehat{n}_\tau \Lambda(\theta^+(k)) > y) + \sum_{k=1}^{n} \mathbb{P}(\widehat{n}_\tau \Lambda(\theta^-(k)) > y).
\end{aligned} \tag{3.B.5}$$

Now we can apply Lemma 3.B.1, with

$$y = v + \sum_{i=1}^{n} \chi^2(P_{S_0}(\cdot \mid z_i), P_{S_{0,\tau,\theta}}(\cdot \mid z_i)) \tag{3.B.6}$$

from which it follows that, for $k = 1, ..., n$,

$$\mathbb{P}(\widehat{n}_\tau \Lambda(\theta^{\pm}(k)) > y) \leq e^{-v/2}. \tag{3.B.7}$$

From (3.B.5) and (3.B.7),

$$\mathbb{P}(\widehat{n}_\tau \mathcal{K}(\widehat{\theta}_\tau, \theta) > y) \leq 2ne^{-v/2}. \tag{3.B.8}$$

This and (3.B.6) yields (3.B.1), thus concluding the proof of Lemma 3.B.2. □

Theorem 3.2.1 follows, if we set $v = 2\ln(n)$ in Lemma 3.B.2.

3.B.2 Verification of Condition C2

Lemma 3.B.3. *Assume that the distribution functions given the covariate z of both the survival and censoring time are in the maximal domain of attraction of the Fréchet law with parameters $\theta(z)$ and $\theta_C(z)$ respectively. Then, for any z,*

$$q_F(\tau \,|\, z) \to \frac{\theta(z)}{\theta_C(z) + \theta(z)} \text{ as } \tau \to \infty.$$

where $\theta(z) > 0$ and $\theta_C(z) > 0$.

Proof. We have for any z,

$$q_F(\tau \,|\, z) = \int_\tau^\infty \frac{S(t \,|\, z)}{S(\tau \,|\, z)} \frac{f_C(t \,|\, z)}{S_C(\tau \,|\, z)} dt \in [0, 1], \quad \tau \geq x_0.$$

Since F and F_C are in the maximal domain of attraction of the Fréchet law, for some $\theta(z) > 0$ and $\theta_C(z) > 0$, we have

$$\frac{S(\tau t \,|\, z)}{S(\tau \,|\, z)} \to t^{-1/\theta(z)}, \text{ as } \tau \to \infty,$$

and

$$\frac{S_C(\tau t \,|\, z)}{S_C(\tau \,|\, z)} \to t^{-1/\theta_C(z)}, \text{ as } \tau \to \infty.$$

Therefore,

$$\lim_{\tau \to \infty} q_F(\tau \,|\, z) = \lim_{\tau \to \infty} \int_\tau^\infty \frac{S(t \,|\, z)}{S(\tau \,|\, z)} \frac{f_C(t \,|\, z)}{S_C(\tau \,|\, z)} dt.$$

By the Lebesgue theorem of dominated convergence,

$$\begin{aligned}\lim_{\tau\to\infty} q_F(\tau \mid z) &= \int_1^\infty t^{-1/\theta(z)} \frac{1}{\theta_C(z)\tau} t^{-1/\theta_C(z)-1} d\tau t \\ &= \frac{1}{\theta_C(z)} \int_1^\infty t^{-\frac{\theta_C(z)+\theta(z)}{\theta(z)\theta_C(z)}-1} dt \\ &= \frac{1}{\theta_C(z)} \frac{\theta(z)\theta_C(z)}{\theta_C(z)+\theta(z)} \\ &= \frac{\theta(z)}{\theta_C(z)+\theta(z)}.\end{aligned}$$

□

3.B.3 Proof of Theorem 3.2.2

We begin with an auxiliary theorem.

Theorem 3.B.4. *Assume condition* (**C2**) *and* (**C3**). *Then, there exists a constant* $c > 0$ *such that,*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{K}(\widehat{\theta}_\tau, \theta) \leq c \frac{\sum_{i=1}^n \chi^2(P_{S_0}(. \mid z_i), P_{S_{0,\tau,\theta}}(. \mid z_i)) + 4\ln(n)}{\sum_{i=1}^n S_C(\tau \mid z_i) S(\tau \mid z_i)}\right) = 1,$$

where $P_{S_0}(. \mid z_i)$ *is the cumulative distribution function of* (3.2.6) *and* $P_{S_{0,\tau,\theta}}(. \mid z_i)$ *is the cumulative distribution function of the joint density of the model* (3.2.3).

Proof. Theorem 3.B.4 follows from Theorem 3.2.1 and the following Lemma 3.B.5. □

Lemma 3.B.5. *Assume condition (**C2**). Then, for every* $\tau \geq x_0$,

$$\mathbb{E}(\widehat{n}_\tau) \geq \sum_{i=1}^n S(\tau \mid z_i) S_C(\tau \mid z_i)(1 - q_0),$$

and

$$\mathbb{P}(\widehat{n}_\tau \leq \mathbb{E}(\widehat{n}_\tau)/2) \leq e^{-\mathbb{E}(\widehat{n}_\tau)/8}.$$

Proof. By the density of the model (3.2.6), we have

$$\mathbb{E}(\widehat{n}_\tau) = \sum_{i=1}^n \int_\tau^\infty f(x \mid z_i) S_C(x \mid z_i) dx,$$

where $\widehat{n}_\tau = \sum_{t_i > \tau} \delta_i$. Therefore,

$$\begin{aligned}
\mathbb{E}(\widehat{n}_\tau) &= \sum_{i=1}^{n} \left([-S(x \mid z_i) S_C(x \mid z_i)]_\tau^\infty - \int_\tau^\infty S(x \mid z_i) f_C(x \mid z_i) dx \right) \\
&= \sum_{i=1}^{n} \left(S(\tau \mid z_i) S_C(\tau \mid z_i) - S(\tau \mid z_i) S_C(\tau \mid z_i) \int_\tau^\infty \frac{S(x \mid z_i)}{S(\tau \mid z_i)} \frac{f_C(x \mid z_i)}{S_C(\tau \mid z_i)} dx \right) \\
&= \sum_{i=1}^{n} S(\tau \mid z_i) S_C(\tau \mid z_i)(1 - q_F(\tau \mid z_i)).
\end{aligned}$$

using $q_F(\tau \mid z_i) \leq q_0$ for any z_i proves the first part of the Lemma.

The second inequality follows from the exponential bound for binomial random variables, and is established by using standard techniques going back to Chernoff [**?**].

□

Lemma 3.B.6. *Assume that Q and Q_0 are two equivalent probability measures on a measurable space. Then,*

$$\chi^2(Q, Q_0) \leq \int \left(\ln \frac{dQ}{dQ_0} \right)^2 \exp \left| \frac{dQ}{dQ_0} \right| dQ.$$

Proof. The proof can be found in the article [89]. □

Now we proceed to prove Theorem 3.2.2. We start by providing the following bound:

$$\max_{z \in \mathbb{Z}} \chi^2(P_{S_0}(\cdot \mid z), P_{S_{0,\tau,\theta}}(\cdot \mid z)) \leq O_{\mathbb{P}} \left(\rho_\tau^2 \max_{z \in \mathbb{Z}} S(\tau \mid z) S_C(\tau \mid z) \right).$$

By Lemma 3.B.6, we have

$$\begin{aligned}
&\max_{z \in \mathbb{Z}} \chi^2(P_{S_0}(. \mid z), P_{S_{0,\tau,\theta}}(. \mid z)) \\
&\leq \max_{z \in \mathbb{Z}} \int \ln^2 \frac{dP_{S_0}(t, \delta \mid z)}{dP_{S_{0,\tau,\theta}}(t, \delta \mid z)} \exp \left| \ln \frac{dP_{S_0}(t, \delta \mid z)}{dP_{S_{0,\tau,\theta}}(t, \delta \mid z)} \right| P_{S_0}(dt, d\delta \mid z).
\end{aligned}$$

For any $x > \tau$, we have

$$\begin{aligned}
\ln \frac{dP_{S_0}(t, \delta \mid z)}{dP_{S_{0,\tau,\theta}}(t, \delta \mid z)} &= \ln \frac{h(t \mid z)^\delta S(t \mid z)}{h_\theta(t \mid z)^\delta S_{\theta,\tau}(t \mid z)} \\
&= \ln \frac{h_0(t)^\delta (e^{\beta \cdot z})^\delta e^{-\int_\tau^t h_0(u) e^{\beta \cdot z} du}}{h_{0,\tau,\theta}(t)^\delta (e^{\beta \cdot z})^\delta e^{-\int_\tau^t h_{0,\tau,\theta}(u) e^{\beta \cdot z} du}} \\
&= \delta \ln \frac{h_0(t)}{(\theta t)^{-1}} - e^{\beta \cdot z} \int_\tau^t (h_0(u) - \frac{1}{\theta u}) du.
\end{aligned}$$

It follows

$$\begin{aligned}\left|\ln\frac{dP_{S_0}(t,\delta\,|\,z)}{dP_{S_{0,\tau,\theta}}(t,\delta\,|\,z)}\right| &= \left|\delta\ln\frac{h_0(t)}{(\theta t)^{-1}} - \int_\tau^t\left(h(u\,|\,z) - \frac{1}{\theta u}e^{\beta\cdot z}\right)du\right| \\ &= \left|\delta\ln(\frac{th_0(t)}{\theta^{-1}} - 1 + 1) - \int_\tau^t\left(uh(u\,|\,z) - \frac{1}{\theta}e^{\beta\cdot z}\right)\frac{du}{u}\right| \\ &\leq \delta 2\theta\left|th_0(t) - \frac{1}{\theta}\right| + \left|\int_\tau^t\left(uh(u\,|\,z) - \frac{1}{\theta}e^{\beta\cdot z}\right)\frac{du}{u}\right|.\end{aligned}$$

Let $\rho_\tau = \sup_{t>\tau}\left|th_0(t) - \frac{1}{\theta}\right|$, then

$$\left|\ln\frac{dP_{S_0}(t,\delta\,|\,z)}{dP_{S_{0,\tau,\theta}}(t,\delta\,|\,z)}\right| \leq \rho_\tau\left(\delta 2\theta + e^{\beta\cdot z}\ln\frac{t}{\tau}\right).$$

We have

$$\begin{aligned}&\max_{z\in\mathbb{Z}}\chi^2(P_{S_0}(\cdot\,|\,z), P_{S_{0,\tau,\theta}}(\cdot\,|\,z)) \\ &\leq \max_{z\in\mathbb{Z}}\int\left(\ln\frac{dP_{S_0}(\cdot\,|\,z)}{dP_{S_{0,\tau,\theta}}(\cdot\,|\,z)}\right)^2\exp\left|\ln\frac{dP_{S_0}(\cdot\,|\,z)}{dP_{S_{0,\tau,\theta}}(\cdot\,|\,z)}\right|dP_{S_0}(\cdot\,|\,z) \\ &\leq \max_{z\in\mathbb{Z}}\int_\tau^\infty\sum_{\delta\in\{0,1\}}\rho_\tau^2\left(\delta 2\theta + e^{\beta\cdot z}\ln\frac{t}{\tau}\right)^2\exp(\rho_\tau\left(\delta 2\theta + e^{\beta\cdot z}\ln\frac{t}{\tau}\right))P_{S_0}(t,\delta\,|\,z)dt \\ &\leq \max_{z\in\mathbb{Z}}\int_\tau^\infty\rho_\tau^2 e^{2\beta\cdot z}\ln^2\frac{t}{\tau}\exp\left(\rho_\tau e^{\beta\cdot z}\ln\frac{t}{\tau}\right)S(t\,|\,z)f_C(t\,|\,z) \\ &+ \rho_\tau^2\left(2\theta + e^{\beta\cdot z}\ln\frac{t}{\tau}\right)^2\exp\left(\rho_\tau(2\theta + e^{\beta\cdot z}\ln\frac{t}{\tau})\right)f(t\,|\,z)S_C(t\,|\,z)dt \\ &\leq \max_{z\in\mathbb{Z}}\int_\tau^\infty\rho_\tau^2\left(\frac{t}{\tau}\right)^{\rho_\tau}\left(e^{2\beta\cdot z}\ln^2\frac{t}{\tau} + \left(2\theta + e^{\beta\cdot z}\ln\frac{t}{\tau}\right)^2 e^{2\rho_\tau\theta}\right) \\ &\qquad\times\left(f(t\,|\,z)S_C(t\,|\,z) + S(t\,|\,z)f_C(t\,|\,z)\right)dt.\end{aligned}$$

Let $g(u) = \left(e^{2\beta\cdot z}u^2 + (2\theta + e^{\beta\cdot z}u)^2 e^{2\rho_\tau\theta}\right)e^{\rho_\tau u}$. Then

$$\begin{aligned}&\max_{z\in\mathbb{Z}}\chi^2(P_{S_0}(.\,|\,z), P_{S_{0,\tau,\theta}}(.\,|\,z)) \\ &\leq \max_{z\in\mathbb{Z}}\int_\tau^\infty\rho_\tau^2 g\left(\ln\frac{t}{\tau}\right)(f(t\,|\,z)S_C(t\,|\,z) + S(t\,|\,z)f_C(t\,|\,z))dt.\end{aligned}$$

Since $S(t\,|\,z) \leq S(\tau\,|\,z)$ and $S_C(t\,|\,z) \leq S_C(\tau\,|\,z)$ for every $t > \tau$, we obtain

$$\begin{aligned}&\max_{z\in\mathbb{Z}}\chi^2(P_{S_0}(.\,|\,z), P_{S_{0,\tau,\theta}}(.\,|\,z)) \\ &\leq \max_{z\in\mathbb{Z}}\rho_\tau^2\int_\tau^\infty g\left(\ln\frac{t}{\tau}\right) \\ &\qquad\times\left(\frac{f(t\,|\,z)}{S(\tau\,|\,z)}S(\tau\,|\,z)S_C(\tau\,|\,z) + \frac{f_C(t\,|\,z)}{S_C(\tau\,|\,z)}S(\tau\,|\,z)S_C(\tau\,|\,z)\right)dt \\ &\leq \max_{z\in\mathbb{Z}}\rho_\tau^2 S(\tau\,|\,z)S_C(\tau\,|\,z)\int_\tau^\infty g\left(\ln\frac{t}{\tau}\right)\left(\frac{f(t\,|\,z)}{S(\tau\,|\,z)} + \frac{f_C(t\,|\,z)}{S_C(\tau\,|\,z)}\right)dt.\end{aligned}$$

We can rewrite the ratio $\frac{S(t\,|\,z)}{S(\tau\,|\,z)}$ as $e^{-\int_\tau^t h(u\,|\,z)du} = e^{-\int_\tau^t uh(u\,|\,z)\frac{du}{u}}$. We know that $th(t\,|\,z)$ is bounded below for t large enough by : $th(t\,|\,z) \geq \frac{e^{\beta\cdot z}}{2\theta}$. Then,

$$\frac{S(t\,|\,z)}{S(\tau\,|\,z)} \leq e^{-\frac{e^{\beta\cdot z}}{2\theta}\ln\frac{t}{\tau}}.$$

We know that :

$$\begin{aligned}
\frac{d}{dt}g\left(\ln\frac{t}{\tau}\right) &= \frac{d}{dt}\left(\ln^2\frac{t}{\tau} + \left(2\theta + \ln\frac{t}{\tau}\right)^2 e^{2\rho_\tau\theta}\right)e^{\rho_\tau\ln\frac{t}{\tau}} \\
&= \left[\frac{2}{t}\ln\frac{t}{\tau} + e^{2\rho_\tau\theta}\frac{2}{t}\left(2\theta + \ln\frac{t}{\tau}\right)\right]e^{\rho_\tau\ln\frac{t}{\tau}} \\
&+ \left(\ln^2\frac{t}{\tau} + \left(2\theta + \ln\frac{t}{\tau}\right)^2 e^{2\rho_\tau\theta}\right)e^{\rho_\tau\ln\frac{t}{\tau}}\rho_\tau\frac{1}{t} \\
&= \left(\frac{t}{\tau}\right)^{\rho_\tau-1}\Big[\frac{2}{\tau}\ln\frac{t}{\tau} + \frac{2}{\tau}e^{2\rho_\tau\theta}2\theta + \frac{2}{\tau}e^{2\rho_\tau\theta}\ln\frac{t}{\tau} + \frac{\rho_\tau}{\tau}\ln^2\frac{t}{\tau} \\
&\quad + \frac{\rho_\tau}{\tau}4\theta^2e^{2\rho_\tau\theta} + \frac{\rho_\tau}{\tau}4\theta\ln\frac{t}{\tau}e^{2\rho_\tau\theta} + \frac{\rho_\tau}{\tau}\ln^2\frac{t}{\tau}e^{2\rho_\tau\theta}\Big].
\end{aligned}$$

Integrating by parts the term $\int_\tau^\infty g\left(\ln\frac{t}{\tau}\right)\left(\frac{f(t\,|\,z)}{S(\tau\,|\,z)}\right)dt$, we have :

$$\begin{aligned}
&\int_\tau^\infty g\left(\ln\frac{t}{\tau}\right)\frac{f(t\,|\,z)}{S(\tau\,|\,z)}dt \\
&= \left[-g\left(\ln\frac{t}{\tau}\right)\frac{S(t\,|\,z)}{S(\tau\,|\,z)}\right]_\tau^\infty + \int_\tau^\infty \frac{S(t\,|\,z)}{S(\tau\,|\,z)}g'\left(\ln\frac{t}{\tau}\right)dx \\
&\leq 4\theta^2e^{2\rho_\tau\theta} + \int_\tau^\infty e^{-\frac{e^{\beta\cdot z}}{2\theta}\ln\frac{t}{\tau}}g'(\ln\frac{t}{\tau})dt \\
&\leq 4\theta^2e^{2\rho_\tau\theta} + \int_\tau^\infty \left(\frac{t}{\tau}\right)^{\rho_\tau-\frac{e^{\beta\cdot z}}{2\theta}-1}\cdot\Big[\frac{2}{\tau}\ln\frac{t}{\tau} + \frac{2}{\tau}e^{2\rho_\tau\theta}2\theta + \frac{2}{\tau}e^{2\rho_\tau\theta}\ln\frac{t}{\tau} \\
&\qquad + \frac{\rho_\tau}{\tau}\ln^2\frac{t}{\tau} + \frac{\rho_\tau}{\tau}4\theta^2e^{2\rho_\tau\theta} + \frac{\rho_\tau}{\tau}4\theta\ln\frac{t}{\tau}e^{2\rho_\tau\theta} + \frac{\rho_\tau}{\tau}\ln^2\frac{t}{\tau}e^{2\rho_\tau\theta}\Big]dt.
\end{aligned}$$

We know that ρ_τ is supposed to be small for large values of t. It is safe to say that $t^{\rho_\tau-\frac{e^{\beta\cdot z}}{2\theta}-1} \leq t^{-\frac{1}{4\theta}}$. Then, $\int_\tau^\infty g\left(\ln\frac{t}{\tau}\right)\frac{f(t\,|\,z)}{S(\tau\,|\,z)}dt$ can be bounded by a constant.

$$\int_\tau^\infty g\left(\ln\frac{t}{\tau}\right)\frac{f(t\,|\,z)}{S(\tau\,|\,z)}dt = O_{\mathbb{P}}(1).$$

In the same way, for $h_C(t\,|\,z) \geq c''$, for $t \geq \tau$, we have

$$\int_\tau^\infty g\left(\ln\frac{t}{\tau}\right)\frac{f_C(t\,|\,z)}{S_C(t\,|\,z)}dt = O_{\mathbb{P}}(1).$$

Then:

$$\max_{z\in\mathbb{Z}}\chi^2(P_{S_0}(\cdot\,|\,z), P_{S_{0,\tau,\theta}}(\cdot\,|\,z)) \leq O_{\mathbb{P}}\left(\rho_\tau^2\max_{z\in\mathbb{Z}}S(\tau\,|\,z)S_C(\tau\,|\,z)\right).$$

Using the previous bound with $\tau = \tau_n$, we have, as $n \to \infty$,

$$\sum_{i=1}^{n} \chi^2(P_{S_0}(\cdot \,|\, z_i), P_{S_{0,\tau_n,\theta}}(\cdot \,|\, z_i)) \leq n\rho_{\tau_n}^2 \max_{z\in\mathbb{Z}} S(\tau_n \,|\, z) S_C(\tau_n \,|\, z) \to 0, \tag{3.B.9}$$

since $\max_{z\in\mathbb{Z}} S(\tau_n \,|\, z) S_C(\tau_n \,|\, z) \leq 1$ and by condition (3.2.13),

$$n\rho_{\tau_n}^2 \to 0 \quad \text{as} \quad n \to \infty.$$

On the other hand, from condition (3.2.14), we have

$$\sum_{i=1}^{n} S_C(\tau_n \,|\, z_i) S(\tau_n \,|\, z_i) \to \infty \quad \text{as} \quad n \to \infty.$$

Since, by Theorem 3.B.4, we have:

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{K}(\widehat{\theta}_{\tau_n}, \theta) \leq c \frac{\sum_{i=1}^{n} \chi^2(P_{S_0}(\cdot \,|\, z_i), P_{S_{0,\tau_n,\theta}}(\cdot \,|\, z_i)) + 4\ln(n)}{\sum_{i=1}^{n} S_C(\tau_n \,|\, z_i) S(\tau_n \,|\, z_i)} \right) = 1,$$

using (3.2.10), it is easy to see that $\mathcal{K}(\widehat{\theta}_{\tau_n}, \theta) \to 0$ as $n \to \infty$, which means that

$$\widehat{\theta}_{\tau_n} \xrightarrow[n\to\infty]{\mathbb{P}} \theta.$$

3.B.4 Proof of Theorem 3.3.1

Proof. Starting from the auxiliary result in the proof for the Theorem 3.2.2, we have

$$\max_{z\in\mathbb{Z}} \chi^2(P_{S_0}(\cdot \,|\, z), P_{S_{0,\tau,\theta}}(\cdot \,|\, z)) \leq O_{\mathbb{P}}\left(\rho_\tau^2 \max_{z\in\mathbb{Z}} S(\tau \,|\, z) S_C(\tau \,|\, z) \right) \quad \text{as } \tau \to \infty.$$

We now want to find a sequence of threshold (τ_n) such that

$$\max_{z\in\mathbb{Z}} S(\tau_n \,|\, z) S_C(\tau_n \,|\, z) \rho_{\tau_n}^2 \leq c_0 \left(\frac{\ln n}{n} \right).$$

Suppose that there exists a sequence of (τ_n) and a constant c_0 such that

$$\max_{z\in\mathbb{Z}} \chi^2(P_{S_0}(. \,|\, z), P_{S_{0,\tau_n,\theta_{\tau_n}}}(. \,|\, z)) = c_0 \frac{\ln n}{n}. \tag{3.B.10}$$

Since the baseline hazard function is assumed to satisfy (**C5**), we have,

$$|xh(x \,|\, z) - \frac{1}{\theta} e^{\beta\cdot z}| \leq c_1' x^{-\frac{\alpha e^{\beta\cdot z}}{\theta}}.$$

From (**C5**), we find the following lower bound for $xh(x \,|\, z)$:

$$\begin{aligned} xh(x \,|\, z) &\geq \frac{e^{\beta\cdot z}}{\theta} - |xh(x \,|\, z) - \frac{e^{\beta\cdot z}}{\theta}| \\ &\geq \frac{e^{\beta\cdot z}}{\theta} - c_1 x^{-\frac{\alpha e^{\beta\cdot z}}{\theta}}. \end{aligned}$$

We note that

$$\begin{aligned}
\max_{z\in\mathbb{Z}} S(\tau_n \mid z) &= \max_{z\in\mathbb{Z}} \exp\left(-\int_{x_0}^{\tau_n} h(t \mid z)dt\right) \\
&= \max_{z\in\mathbb{Z}} \exp\left(-\int_{x_0}^{\tau_n} th(t \mid z)\frac{dt}{t}\right) \\
&\leq \max_{z\in\mathbb{Z}} \exp\left(-\frac{e^{\beta\cdot z}}{\theta}\ln\frac{\tau_n}{x_0} - \frac{\theta c_1'}{\alpha e^{\beta\cdot z}}\left(\tau_n^{-\frac{\alpha e^{\beta\cdot z}}{\theta}} - x_0^{-\frac{\alpha e^{\beta\cdot z}}{\theta}}\right)\right) \\
&\leq \max_{z\in\mathbb{Z}} \exp\left(-\frac{e^{\beta\cdot z}}{\theta}\ln\frac{\tau_n}{x_0}\right)\exp\left(\frac{c_1'}{\alpha}x_0^{-\alpha}\right).
\end{aligned}$$

We now find bounds for $xh_C(x \mid z)$.

$$\begin{aligned}
xh_C(x \mid z) - \frac{\gamma e^{\beta\cdot z}}{\theta} &\geq -|xh_C(x \mid z) - \frac{\gamma e^{\beta\cdot z}}{\theta}| \\
xh_C(x \mid z) &\geq \frac{\gamma e^{\beta\cdot z}}{\theta} - |xh_C(x \mid z) - \frac{\gamma e^{\beta\cdot z}}{\theta}| \\
&\geq \frac{\gamma e^{\beta\cdot z}}{\theta} - c_2 x^{-\mu}.
\end{aligned}$$

We have

$$\begin{aligned}
\max_{z\in\mathbb{Z}} S_C(\tau_n \mid z) &= \max_{z\in\mathbb{Z}} \exp\left(-\int_{x_0}^{\tau_n} h_C(x \mid z)dx\right) \\
&= \max_{z\in\mathbb{Z}} \exp\left(-\int_{x_0}^{\tau_n} xh_C(x \mid z)\frac{dx}{x}\right) \\
&\leq \max_{z\in\mathbb{Z}} \exp\left(-\frac{\gamma e^{\beta\cdot z}}{\theta}\ln\frac{\tau_n}{x_0} - \frac{c_2}{\mu}\left(\tau_n^{-\mu} - x_0^{-\mu}\right)\right) \\
&\leq \max_{z\in\mathbb{Z}} \exp\left(-\frac{\gamma e^{\beta\cdot z}}{\theta}\ln\frac{\tau_n}{x_0}\right)\exp\left(\frac{c_2}{\mu}x_0^{-\mu}\right).
\end{aligned}$$

Then,

$$\max_{z\in\mathbb{Z}} S(\tau_n \mid z)S_C(\tau_n \mid z)\rho_{\tau_n}^2 \leq \max_{z\in\mathbb{Z}}(c_1')^2\tau_n^{-(\frac{e^{\beta\cdot z}}{\theta}+\frac{\gamma e^{\beta\cdot z}}{\theta}+2\frac{\alpha e^{\beta\cdot z}}{\theta})}x_0^{\frac{e^{\beta\cdot z}}{\theta}+\frac{\gamma e^{\beta\cdot z}}{\theta}}c_3. \quad (3.B.11)$$

Solving the equation (3.B.11) for τ_n yields

$$\begin{aligned}
&\max_{z\in\mathbb{Z}}(c_1')^2\tau_n^{-(\frac{e^{\beta\cdot z}}{\theta}+\frac{\gamma e^{\beta\cdot z}}{\theta}+2\frac{\alpha e^{\beta\cdot z}}{\theta})}x_0^{\frac{e^{\beta\cdot z}}{\theta}+\frac{\gamma e^{\beta\cdot z}}{\theta}}c_3 = c_0\left(\frac{\ln n}{n}\right), \\
&\tau_n = n^{\frac{\theta}{\min_{z\in Z}(e^{\beta\cdot z})(1+\gamma+2\alpha)}}\ln^{-\frac{\theta}{\min_{z\in Z}(e^{\beta\cdot z})(1+\gamma+2\alpha)}} n, \\
&\tau_n = n^{\frac{\theta/\min_{z\in Z}(e^{\beta\cdot z})}{1+\gamma+2\alpha}}\ln^{-\frac{\theta/\min_{z\in Z}(e^{\beta\cdot z})}{1+\gamma+2\alpha}} n.
\end{aligned}$$

Now, we search a lower bound for $\sum_{i=1}^{n} S_(\ \tau \,|\, z_i) S_C(\tau_n \,|\, z_i)$. We have for any z,

$$\begin{aligned} S(\tau_n \,|\, z) &= \exp\left(-\int_{x_0}^{\tau_n} th(t \,|\, z)\frac{dt}{t}\right) \\ &\geq \exp\left(-\frac{e^{\beta\cdot z}}{\theta}\ln\frac{\tau_n}{x_0} - \frac{\theta c_1}{-\alpha e^{\beta\cdot z}}\left(\tau_n^{-\frac{\alpha e^{\beta\cdot z}}{\theta}} - x_0^{-\frac{\alpha e^{\beta\cdot z}}{\theta}}\right)\right) \\ &\geq \exp\left(-\frac{e^{\beta\cdot z}}{\theta}\ln\frac{\tau_n}{x_0}\right)\exp\left(-\frac{\theta c_1'}{\alpha e^{\beta\cdot z}}x_0^{-\frac{\alpha e^{\beta\cdot z}}{\theta}}\right), \end{aligned}$$

and

$$\begin{aligned} S_C(\tau_n \,|\, z) &= \exp\left(-\int_{x_0}^{\tau_n} x h_C(x)\frac{dx}{x}\right) \\ &\geq \exp\left(-\frac{\gamma e^{\beta\cdot z}}{\theta}\ln\frac{\tau_n}{x_0}\right)\exp\left(-\frac{c_2}{\mu}x_0^{-\mu}\right). \end{aligned}$$

Choosing z such as minimizing $\sum_{i=1}^{n} S_(\ \tau \,|\, z_i) S_C(\tau_n \,|\, z_i)$ yields the result of the theorem. □

Chapitre 4

Application sur données électriques

L'objectif principal de ce chapitre est de pouvoir analyser les données de consommations électriques individuelles entre toute la période d'étude, soit décembre 2015 à mars 2018, et de comparer les résultats avec les différentes données dont nous disposons (panel d'appartenance, enquêtes).

Nous donnons ici un rappel des interrogations auxquelles nous allons essayer de répondre dans ce chapitre.

- Peut-on définir des foyers à risque lors des écrêtements ?
- Les accompagnements individuels ou collectifs ont-ils un impact sur la consommation électrique ?

Nous commencerons dans la Section 4.2 par une étude sur les écrêtements et les foyers à risque et nous finirons en Section 4.3 par étudier l'impact des visites individuelles sur la consommation électrique.

Avant d'étudier les écrêtements et l'impact des visites individuelles sur la consommation électrique, nous réalisons une classification des foyers selon leur type de chauffage.

4.1 Mode de chauffage des foyers

Parmi les données recueillies lors des enquêtes, nous avons l'information sur le mode de chauffage principal des foyers. Celle-ci est primordiale car l'effet de la température

extérieure et le mode de consommation des foyers n'est pas le même pour les foyers avec un chauffage électrique que ceux avec un autre mode de chauffage (bois, gaz, ...). Nous avons l'information du type de chauffage pour 375 foyers, mais nous voulons cette information pour l'ensemble des foyers sur lesquels nous avons des données de courbes de charge. Pour prédire le mode de chauffage de chaque foyer, nous nous sommes appuyés sur la méthode des forêts aléatoires pour classer les foyers en deux groupes :

- Les foyers ayant un mode de chauffage principal électrique : CE
- Les foyers ayant un autre mode de chauffage principal : AC

Nous allons utiliser plusieurs mois des hivers 2016-2017 et 2017-2018 pour avoir une meilleure précision sur notre classification. La démarche suivie est la suivante : pour chaque mois des deux hivers (le premier hiver est écarté dû au grand nombre de données manquantes), nous allons prédire le mode de chauffage principal des foyers dont nous ne disposons pas de l'information. Les résultats sont regroupés pour donner une probabilité d'appartenance à un groupe (chauffage électrique ou autre). Enfin, le mode de chauffage principal d'un foyer va être sélectionné en gardant la probabilité d'appartenance la plus élevée. Si celle-ci est trop proche de 50%, nous allons considérer ce foyer comme non classifiable. Illustrons par un exemple sur deux mois :

- si un foyer a été classé comme ayant un chauffage électrique principal lors des deux mois, alors nous allons le définir comme ayant un mode de chauffage principal électrique.
- si un foyer a été classé comme ayant un chauffage électrique principal lors d'un mois et que l'autre mois il a été classé comme n'ayant pas de mode de chauffage principal électrique, alors nous ne pouvons pas définir le mode de chauffage pour ce foyer.
- si un foyer a été classé comme ayant un autre mode de chauffage lors des deux mois, alors nous allons le définir comme ayant un mode de chauffage principal autre.

Nous n'allons regarder que les foyers 6 kVA et 9 kVA pour considérer le mode de chauffage principal. Les foyers avec une puissance souscrite de 3 kVA n'ont généralement pas de chauffage principal électrique alors que ceux avec une puissance souscrite de 12 kVA ont en général tous un chauffage principal électrique.

Pour les foyers avec un contrat 6 kVA, nous avons 183 foyers dont nous connaissons le mode de chauffage principal et 325 foyers qui n'ont pas répondu au questionnaire. Comme nous avons un nombre correct de foyers dont nous avons l'information

sur le chauffage principal, nous pouvons utiliser ces informations pour tester la méthode utilisée. C'est à dire que parmi les foyers que nous connaissons, nous allons choisir au hasard un échantillon d'apprentissage sur lequel la méthode va être utilisée et un échantillon où nous allons tester la méthode dans le but de regarder l'erreur que nous allons avoir. Sur 1000 échantillonnages des données dont nous disposons le mode de chauffage principal, en moyenne, nous avons 7% d'erreur de classification sur chaque mois. Nous avons décidé de ne pas classifier les foyers dont la probabilité d'appartenance à une classe se trouve entre 25% et 75%.

Moyen de consommation pour le chauffage	**Échantillon d'apprentissage**	**Échantillon test**
Électrique	23	26
Autre	165	280
Ne peut pas classifier	0	19

TABLE 4.1 : Analyse discriminante pour les foyers avec un contrat d'une puissance souscrite de 6 kVA.

La Table 4.1 présente les résultats de la classification pour les foyers avec un contrat d'une puissance souscrite de 6 kVA. Les foyers que nous n'avons pas classés ne vont pas être inclus dans les études où nous nous servons de cette information, soit la Section 4.2.

Pour les foyers 9 kVA, nous avons 221 foyers avec assez de données en janvier 2017 pour réaliser l'analyse. De plus, nous avons l'information sur le mode de chauffage principal pour 72 d'entre eux mais nous n'avons pas cette information pour 149 foyers. Nous avons en moyenne une erreur de classification de 17% (ce qui correspond à 2-3 foyers quand nous utilisons 80% des foyers en tant qu'échantillon d'apprentissage) sur chaque mois. Cette erreur est due au faible nombre de foyers dont nous disposons de l'information du mode de chauffage principal.

Moyen de consommation pour le chauffage	**Échantillon d'apprentissage**	**Échantillon test**
Électrique	47	95
Autre	25	38
Ne peut pas classifier	0	16

TABLE 4.2 : Analyse discriminante pour les foyers avec un contrat d'une puissance souscrite de 9 kVA.

La Table 4.2 présente le résultat de l'analyse pour les foyers avec un contrat d'une puissance souscrite de 9 kVA. On remarque que nous avons 16 foyers qui ne

sont pas classés, cela représente 10% des foyers sans l'information sur le mode de chauffage. De façon similaire aux foyers avec un contrat 6 kVA, nous n'allons pas utiliser ces foyers lors des études Section 4.2.

Pour l'utilisation de l'information sur le mode de chauffage, nous allons définir les abréviations suivantes :

- CE : Foyers considérés avec un mode de chauffage principal électrique.
- AC : Foyers considérés avec un autre mode de chauffage principal.

4.2 Écrêtements

Cette section va être séparée en deux parties, dans un premier temps, nous allons estimer les probabilités de dépasser la puissance maximale des foyers pendant les périodes d'écrêtements. Ensuite, nous allons nous placer du point de vue des expérimentateurs et proposer une réduction de la puissance maximale gênant le moins possible les expérimentateurs. La méthode décrite dans le Chapitre 3 va être utilisée dans cette section. Les covariables prises en compte sont l'heure de la mesure, le mois, le jour de la semaine et la température extérieure.

Les différents panels utilisés dans cette section sont :

- E : le panel témoin.
- C : le panel suivi collectivement.
- IA : le panel suivi individuellement par ALOEN.
- IV : le panel suivi individuellement par Vity.
- ID : le panel suivi individuellement par Delta Dore.

4.2.1 Exemple d'application sur un foyer

Dans cette section, nous détaillons l'utilisation du modèle décrit dans le Chapitre 3 sur un foyer pris au hasard.

Choix des covariables

Le choix des covariables est important pour avoir une meilleure estimation de la consommation électrique du foyer ainsi que les probabilités de dépasser la puissance

maximale en période d'écrêtement. Pour choisir quelles variables garder, nous allons utiliser les critères de sélection AIC (Akkaike Infomation Criterion) et BIC (Bayesian Information Criterion).

Les modèles testés seront ceux avec les variables suivantes :

1. La température extérieure et l'heure.
2. La température extérieure, l'heure et le jour de la semaine.
3. La température extérieure, l'heure, le jour de la semaine et le mois.

Le but étant de définir des foyers à risque, nous n'avons utiliser que la période hivernale pour nos données (début novembre - fin mars). En prenant en compte le deuxième hiver pour référence et pour chacun des foyers, nous avons testé chacun des modèles et retiré les critères de sélection susnommés. Le tableau suivant montre la répartition des foyers selon le modèle ayant le plus faible BIC et AIC. La moyenne des pseudo-R^2 (Cox et Snell [96]) des modèles est également donnée pour information.

	Modèle 1	Modèle 2	Modèle 3
AIC	7	13	779
BIC	34	40	725
Pseudo-R^2	0.296	0.301	0.322

TABLE 4.3 : Nombre de foyers selon le modèle ayant le plus faible AIC et BIC. La moyenne des pseudo-R^2 de Cox et Snell est donnée.

Pour chacun des critères, le modèle ayant le plus faible BIC et AIC est celui prenant en compte la température extérieure, l'heure, le jour de la semaine et le mois. C'est ce que nous avons choisi par la suite.

Proportionnalité des hasards

Le modèle de Cox [6] assume la proportionnalité des fonctions de hasard en fonction des covariables. Bien que celle-ci ne peut pas être vérifier dans notre cas dû au nombre de covariables dont nous disposons, nous avons tout de même regardé comment les fonctions de hasards se comportent en fonction des covariables.

La fonction `cox.zph` du langage R permet de tester par un test du χ^2 la proportionnalité des variables du modèle et donne un visuel sur le paramètre β pour chacune des variables en fonction des données de consommation électrique. Si $\beta(t)$ est constant en fonction des données, alors la proportionnalité des hasards n'est pas

rejetée pour cette variable. Pour plus de détails sur le test effectué, nous référons à Grambsch and Therneau [97].

Au total, nous avons 32 variables dans le modèle, ce qui est conséquent. La Figure 4.1 montre le paramètre $\beta(t)$ pour deux variables. Pour le foyer, nous n'avons qu'une seule variable parmi les 32 dont le test de proportionnalité n'a pas conclut au rejet de celle-ci, il s'agit de la variable de la température extérieure comprise entre 5 et 10 degrés (à droite sur la figure). On peut remarquer que le paramètre β est constant en fonction des données pour cette variable. En revanche, pour les 31 autre variables, la proportionnalité a été rejetée et la représentation des β en fonction des données est similaire au graphe de gauche sur la Figure 4.1.

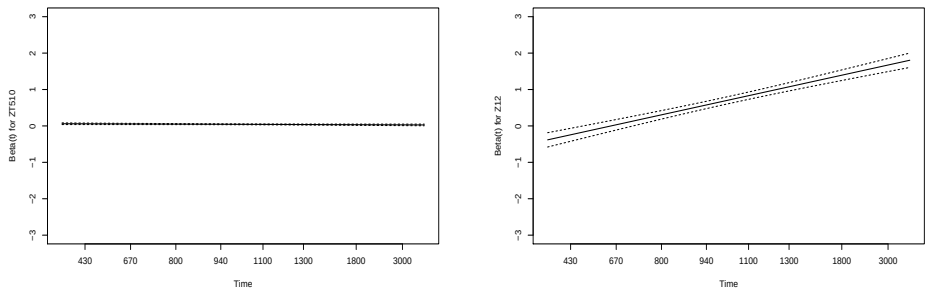

FIGURE 4.1 : β en fonction du temps pour la variable de la température extérieure comprise entre 5 et 10 degrés (gauche) et pour l'heure entre 12h et 13h (droite).

Nous avons décidé d'améliorer la non-proportionnalité des hasards en séparant les données par les tranches horaires suivantes :

- la nuit : entre 22h et 6h
- le matin : entre 6h et 10h
- le midi : entre 10h et 14h
- l'après-midi : entre 14h et 18h
- le soir : entre 18h et 22h.

Lorsque nous réalisons de nouveau le test de proportionnalité pour ce foyer, nous obtenons 6 ou 7 variables dont le test de la proportionnalité n'a pas conduit au rejet de celle-ci par tranche horaire. Cette séparation a été utilisée dans la Section 4.2.2 et 4.2.3.

Application sur le foyer

Pour le foyer, nous avons estimé un modèle pour chaque tranche horaire définie ci-dessus. Pour chacune de celles-ci, nous avons une fonction de survie baseline (3.2.8).

- $\widehat{S}_{0,\tau,\widehat{\theta}_\tau,\text{nuit}}(t)$
- $\widehat{S}_{0,\tau,\widehat{\theta}_\tau,\text{matin}}(t)$
- $\widehat{S}_{0,\tau,\widehat{\theta}_\tau,\text{midi}}(t)$
- $\widehat{S}_{0,\tau,\widehat{\theta}_\tau,\text{a-m}}(t)$
- $\widehat{S}_{0,\tau,\widehat{\theta}_\tau,\text{soir}}(t)$

À la suite de cette estimation, nous pouvons calculer les probabilités de dépasser une puissance maximale donnée au préalable et avec les covariables correspondantes. Par exemple, si nous voulons calculer la probabilité de dépasser la puissance d'écrêtement lors du premier écrêtement de l'hiver 2015/2016, les covariables sont Z_1 et Z_2 pour respectivement la première et deuxième heure de l'écrêtement. La probabilité d'avoir au moins une mesure dépassant la puissance d'écrêtement pendant le premier écrêtement est :

$$P(\text{Une mesure} > P_{\max}) = 1-(1-\widehat{S}_{0,\tau,\widehat{\theta}_\tau,\text{a-m}}(P_{max})^{e^{\beta\text{a-m}Z_1}})(1-\widehat{S}_{0,\tau,\widehat{\theta}_\tau,\text{soir}}(P_{max})^{e^{\beta\text{soir}}Z_2}).$$

4.2.2 Foyers à risque

Nous avons séparé les foyers expérimentateurs par rapport à leur puissance maximale souscrite. Afin d'estimer la probabilité de dépasser la puissance maximale en période d'écrêtement, la méthode présentée dans le Chapitre 3 a été utilisée sur chacun des foyers. Un foyer est considéré à risque si la probabilité estimée de dépasser la puissance maximale est plus grande qu'une certaine valeur. La probabilité estimée est fonction de la durée d'écrêtement.

Période d'écrêtement	Probabilité pour considérer un foyer à risque
2 heures	0.08
4 heures	0.04

TABLE 4.4 : Probabilité à dépasser pour considérer un foyer à risque selon la durée de l'écrêtement.

Pour une période de deux heures, la probabilité d'avoir une mesure supérieure à la puissance maximale autorisée est de $1/12 \simeq 0.08$ et pour une période de quatre

heures, cette probabilité est de $1/24 \simeq 0.04$. Ces probabilités sont dues au fait que nous avons une mesure toutes les 10 minutes.

Hiver 1

Pour le premier hiver, nous n'avons pas de foyers 3 kVA écrêtés dont nous disposons des données. La Table 4.5 donne le récapitulatif des écrêtements ayant eu lieu lors du premier hiver.

Période d'écrêtement	**Diminution (talon)**	**3 kVA**	**6 kVA**	**9 kVA**	**12 kVA**
11/1/2016 (17-19h)	30% (2.8 kVA)	2.8	4.2	6.3	8.4
13/1/2016 (14-18h)	50% (2.8 kVA)	2.8	3	4.5	6
18/1/2016 (14-18h)	50% (2.8 kVA)	2.8	3	4.5	6
25/2/2016 (14-18h)	50% (2.8 kVA)	2.8	3	4.5	6
1/3/2016 (17-19h)	60% (2.8 kVA)	2.8	2.8	3.6	4.8
7/3/2016 (14-18h)	70% (2.8 kVA)	2.8	2.8	2.8	3.6
18/3/2016 (10-12h)	80% (2.8 kVA)	2.8	2.8	2.8	2.8

TABLE 4.5 : Récapitulatif des écrêtements et de la puissance maximale (en kVA) durant la période d'écrêtement lors de l'hiver 2015/2016.

6 kVA

Pour les foyers 6 kVA CE, nous n'avons que deux foyers respectant un volume de données suffisant. Un foyer a été estimé à risque et seulement pendant le dernier écrêtement. Alors que, il n'y a eu qu'une seule ouverture de breaker sur le foyer n'ayant pas été estimé à risque.

Pour les foyers 6 kVA AC, les résultats sont présentés sur la Figure 4.2. Nous avons 75 foyers durant le premier hiver. On remarque que nous estimons plus de foyers à risque qu'il n'y a de foyer avec au moins une ouverture de breaker.

9 kVA

Pour les 24 foyers avec un contrat d'une puissance souscrite de 9 kVA CE, la Figure 4.3 compare le nombre de foyers estimés à risque avec ceux ayant subi au moins une ouverture de breaker pendant le premier hiver.

On remarque que nous estimons plus de foyers à risque qu'il n'y a eu d'ouvertures de breaker. Les foyers (puissance souscrite 9 kVA) AC sont très peu nombreux (13) pour le premier hiver. Nous avons tout de même un foyer que nous avons considéré à

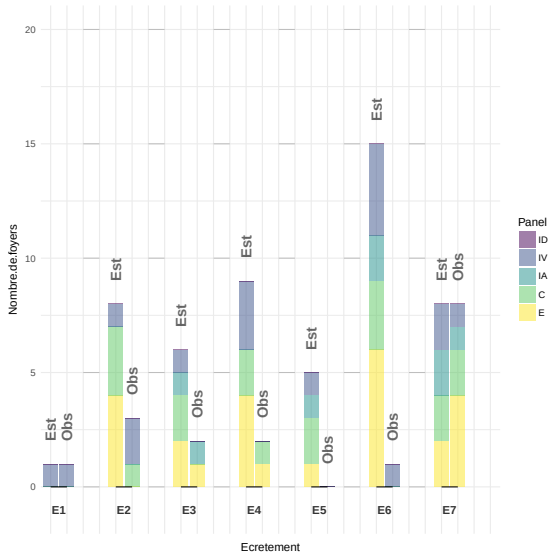

FIGURE 4.2 : Nombre de foyers ayant un contrat 6 kVA AC estimé à risque (Est) et ayant subi au moins une ouverture de breaker (Obs) selon le jour de l'écrêtement et pour chaque panel.

risque pour quasiment tous les écrêtements et ce foyer a subi au moins une ouverture de breaker sur la plupart de ces écrêtements.

12 kVA

Les foyers ayant une puissance souscrite de 12 kVA sont au nombre de deux pour le premier hiver et ont été regroupés dans la table suivante :

Puissance souscrite / Panel	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**	**Écr. 5**	**Écr. 6**	**Écr. 7**
12 kVA / IV	0	0.023	0.025	0.014	0.015	0.198	0.435
12 kVA / IV	0	0.001	0.001	0	0	0.065	0.259

TABLE 4.6 : Probabilité estimée de dépasser la puissance maximale en période d'écrêtement pour les foyers avec une puissance souscrite de 12 kVA pour l'hiver 2015/2016.

Les cases surlignées en rouge dans la Table 4.6 correspondent aux ouvertures de breaker observées durant les écrêtements. Dans ce cas, on peut voir que le foyer à risque avec la plus grande probabilité estimée est celui qui a subi les ouvertures de breaker.

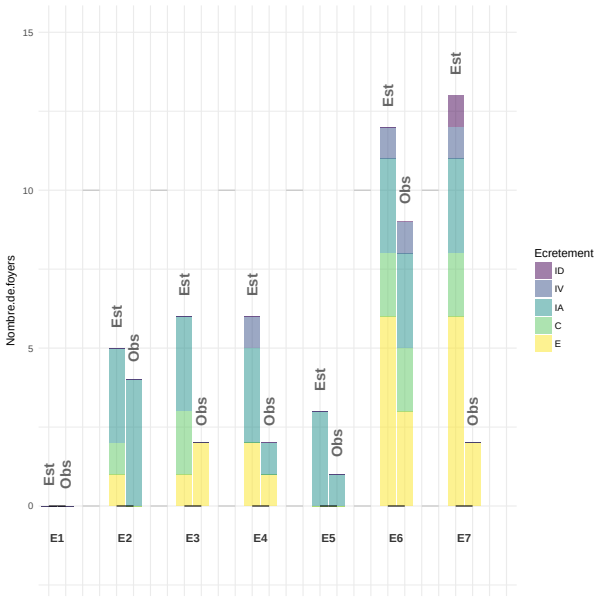

FIGURE 4.3 : Nombre de foyers ayant un contrat 9 kVA CE estimé à risque (Est) et ayant subis au moins une ouverture de breaker (Obs) selon le jour de l'écrêtement et pour chaque panel.

Hiver 2

Le récapitulatif des tirs d'écrêtements ayant eu lieu le second hiver est donné Table 4.7.

Période d'écrêtement	**Diminution (talon)**	**3 kVA**	**6 kVA**	**9 kVA**	**12 kVA**
22/11/2016 (17-21h)	60% (1.5 kVA)	1.5	2.4	3.6	4.8
14/12/2016(14-18h)	70% (1.5 kVA)	1.5	1.8	2.7	3.6
5/1/2017 (6-10h)	50% (2.8 kVA)	2.8	3	4.5	6
17/1/2017 (17-19h)	80% (2 kVA)	2	2	2	2.4
1/2/2017 (14-16h)	80% (1.8 kVA)	1.8	1.8	1.8	2.4
28/2/2017 (18-20h)	80% (1.8 kVA)	1.8	1.8	1.8	2.4
30/3/2017 (7-9h)	80% (1.8 kVA)	1.8	1.8	1.8	2.4

TABLE 4.7 : Récapitulatif des écrêtements et de la puissance maximale (en kVA) durant la période d'écrêtement lors de l'hiver 2016/2017.

3/6 kVA

Il n'y a pas eu d'ouverture de breaker lors de l'hiver 2016/2017 pour les foyers avec un contrat d'une puissance souscrite de 3 kVA. La Table 4.8 donne les probabilités estimées de dépasser la puissance maximale en période d'écrêtement pour chacun des foyers. Les foyers avec une puissance souscrite de 6 kVA CE ne sont pas nombreux. La Table 4.8 permet d'avoir un regard global sur les probabilités estimées de dépasser la puissance maximale ainsi que les ouvertures de breaker que certains foyers ont subi (surlignés en rouge).

Puis. sous. / Panel	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**	**Écr. 5**	**Écr. 6**	**Écr. 7**
3 kVA / E	0.12	0.055	0.057	0.052	0.007	0.025	0.039
3 kVA / C	0	0.005	0	0.019	0.004	0.05	0
3 kVA / C	0	0.043	0.025	0.04	0.004	0.064	0.125
3 kVA / C	0	0.669	0.306	0.028	0.058	0	0.029
6 kVA / E	0.001	0	0	0	0	0	0
6 kVA / C	0.083	0	0.055	0.005	0.013	0.078	
6 kVA / C		0.503	0.758	0.325	0.398	0.609	0.836
6 kVA / C					0.237	0.689	0.613
6 kVA / C		0.09	0.549	0.11	0.001	0.372	0.757
6 kVA / IA	0	0	0	0	0	0.01	0.002
6 kVA / IA	0.589	0	0.005	0.002	0	0.439	0.397
6 kVA / IA	0.001	0	0	0	0.003	0.077	0.05
6 kVA / ID	0.227	0.057	0.32	0.016	0.012	0.177	0.626

TABLE 4.8 : Probabilité estimée de dépasser la puissance maximale en période d'écrêtement pour les foyers avec une puissance souscrite de 3 et 6 kVA CE pour l'hiver 2016/2017.

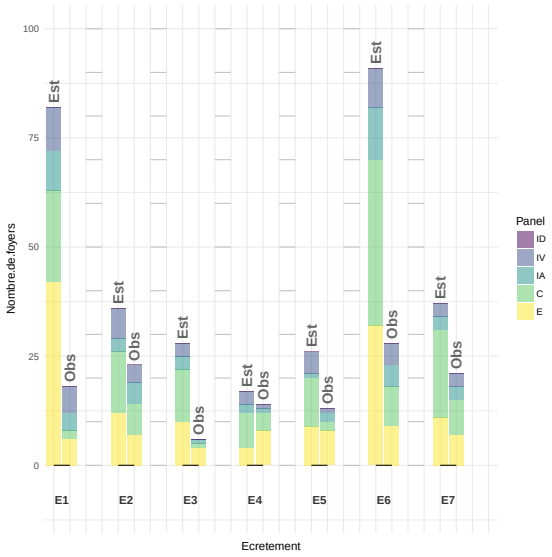

FIGURE 4.4 : Nombre de foyers ayant un contrat 6 kVA AC estimé à risque (Est) et ayant subi au moins une ouverture de breaker (Obs) selon le jour de l'écrêtement et pour chaque panel.

Les cases vides correspondent à des valeurs que nous n'avons pas pu estimer (valeurs de covariables manquantes par exemple). Les foyers avec une puissance souscrite de 6 kVA AC sont au nombre de 156. La Figure 4.4 présente le nombre de foyers avec une probabilité estimée supérieure à 1/12 pour une période d'écrêtement de 2 heures et 1/24 pour une période de 4 heures. Le nombre de foyers ayant subi au moins une ouverture de breaker pendant les écrêtements est également affiché.

9 kVA

Les 52 foyers 9 kVA CE et ayant été estimés à risque sont affichés sur la Figure 4.5 ainsi que ceux ayant subis au moins une ouverture de breaker.

Les foyers avec une puissance souscrite de 9 kVA AC sont au nombre de 26 et la Figure 4.6 présente ceux ayant été estimés à risque et ceux ayant eu une ouverture de breaker pendant les écrêtements.

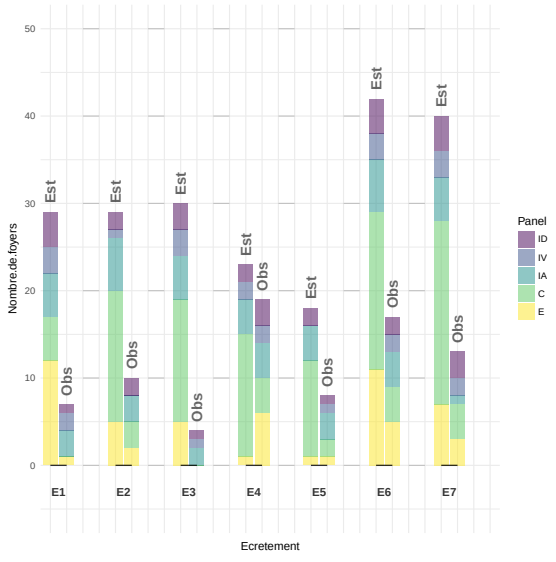

FIGURE 4.5 : Nombre de foyers ayant un contrat 9 kVA CE estimé à risque (Est) et ayant subi au moins une ouverture de breaker (Obs) selon le jour de l'écrêtement et pour chaque panel.

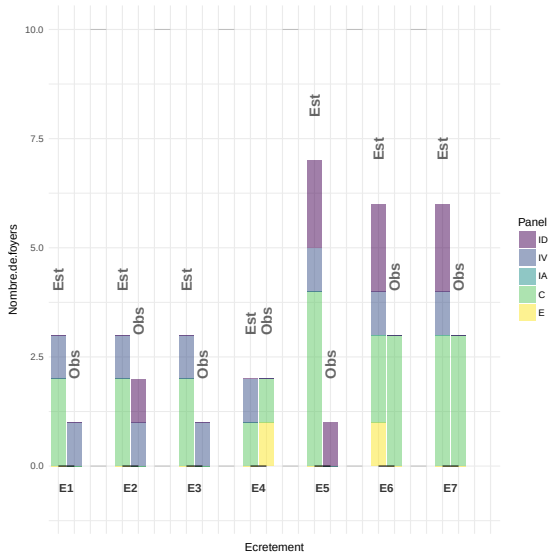

FIGURE 4.6 : Nombre de foyers ayant un contrat 9 kVA AC estimé à risque (Est) et ayant subi au moins une ouverture de breaker (Obs) selon le jour de l'écrêtement et pour chaque panel.

12 kVA

La table suivante donne les probabilités de dépasser la puissance maximale en période d'écrêtement pour les foyers 12 kVA.

Panel	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**	**Écr. 5**	**Écr. 6**	**Écr. 7**
C		0.111	0.034	0.0866	0	0.035	0.037
C					0.001	0.353	0.241
C		0.731	0.757	0.734	0.208	0.805	0.637
C		0.308	0.069	0.362	0.027	0.595	0.05
C		0.034	0.003	0.001	0	0.023	0.026
C					0	0.204	0.114
C		0.112	0.014	0.089	0.007	0.393	0.354
C		0.828	0.981	0.561	0.519	0.891	0.994
C		0.133	0.027	0.031	0.012	0.293	0.662
IV	0.444	0.254	0.518	0.152	0.114	0.401	0.606
IV	0.013	0.001	0.001	0	0	0.03	0.063
ID	0.172	0.062	0.645	0.056	0.142	0.509	0.844

TABLE 4.9 : Probabilités estimées de dépasser la puissance maximale en période d'écrêtement pour les foyers avec une puissance souscrite de 12 kVA pour l'hiver 2016/2017.

Les cases vides correspondent à des valeurs que nous n'avons pas pu estimer (valeurs de covariables manquantes par exemple).

Hiver 3

Pour le troisième hiver, nous avons huit tirs d'écrêtements. Ces tirs sont présentés Table 4.10

Période d'écrêtement	**Diminution (talon)**	**3 kVA**	**6 kVA**	**9 kVA**	**12 kVA**
9/11/2017 (18-20h)	70% (1.8 kVA)	1.8	1.8	2.7	3.6
29/11/2017(18-20h)	80% (1.8 kVA)	1.8	1.8	1.8	2.4
19/12/2017 (17-19h)	80% (2 kVA)	2	2	2	2.4
8/1/2018 (18-20h)	80% (2 kVA)	2	2	2	2.4
2/2/2018 (11-13h)	80% (2 kVA)	2	2	2	2.4
20/2/2018 (18-20h)	80% (2 kVA)	2	2	2	2.4
13/3/2018 (18-20h)	70% (2 kVA)	2	2	2.7	3.6
28/3/2018 (18-20h)	81% (2 kVA)	2	2	2	2.376

TABLE 4.10 : Récapitulatif des écrêtements et de la puissance maximale (en kVA) durant la période d'écrêtement.

3/6 kVA

La Table 4.11 présente les probabilités estimées pour les foyers avec une puissance de 3 kVA et de 6 kVA CE. Les valeurs surlignées en rouge correspondent aux ouvertures de breaker ayant eu lieu. La Figure 4.7 montre le nombre de foyers considérés à risque lors de l'hiver 3 parmi les 158 foyers avec un contrat 6 kVA AC (Est). Le nombre de foyers ayant subis au moins une ouverture de breaker est également affiché (Obs).

9 kVA

Les foyers avec un contrat d'une puissance souscrite de 9 kVA estimés à risque sont présentés sur la Figure 4.8 pour les 53 foyers CE et sur la Figure 4.9 pour les 27 autres.

Puissance souscrite / Panel	Écr. 1	Écr. 2	Écr. 3	Écr. 4	Écr. 5	Écr. 6	Écr. 7	Écr. 8
3 kVA / E	0.023	0.024	0.073	0.042	0	0	0.018	0.019
3 kVA / C	0.01	0.01	0.006	0.064	0	0.013	0.057	0.058
3 kVA / C	0.038	0.044	0.013	0.043	0.053	0.032	0.029	0.034
3 kVA / C	0	0	0.002	0.116	0	0	0	0
6 kVA / E	0.001	0.002	0.007	0.011	0.11	0.002	0.002	0.003
6 kVA / C	0.009	0.011	0.004	0.018	0.009	0.008	0.004	0.006
6 kVA / C	0.598	0.673	0.836	0.816	0.959	0.528	0.397	0.481
6 kVA / C	0.76	0.807	0.765	0.868	0.459	0.696	0.623	0.687
6 kVA / C	0.097	0.148	0.085	0.309	0.096	0.071	0.035	0.061
6 kVA / IA	0	0	0	0.001	0	0	0	0
6 kVA / IA	0.001	0.003	0.037	0.322	0.131	0.18	0.055	0.087
6 kVA / IA	0.107	0.123	0.089	0.128	0	0.109	0.041	0.049
6 kVA / ID	0.355	0.432	0.412	0.56	0.678	0.231	0.156	0.215

TABLE 4.11 : Probabilités estimées de dépasser la puissance maximale en période d'écrêtement pour les foyers avec une puissance souscrite de 3 et 6 kVA CE pour l'hiver 2017/2018.

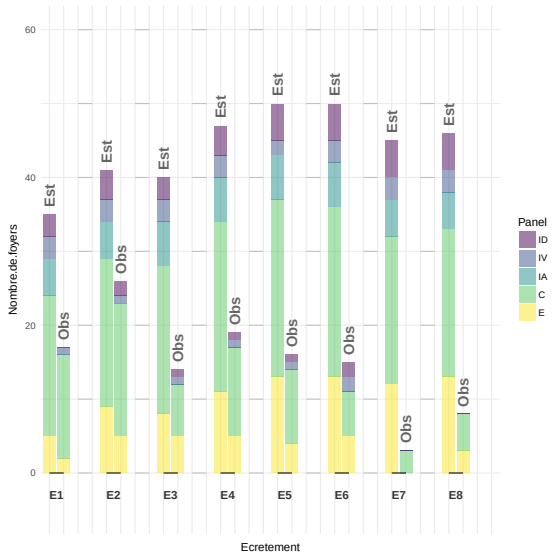

FIGURE 4.8 : Nombre de foyers ayant un contrat 9 kVA CE estimé à risque (Est) et ayant subi au moins une ouverture de breaker (Obs) selon le jour de l'écrêtement et pour chaque panel.

12 kVA

Les probabilités estimées de dépasser la puissance maximale en période d'écrêtement pour les foyers 12 kVA sont présentées Table 4.12. Les probabilités surlignées en rouge correspondent aux ouvertures de breaker ayant eu lieu.

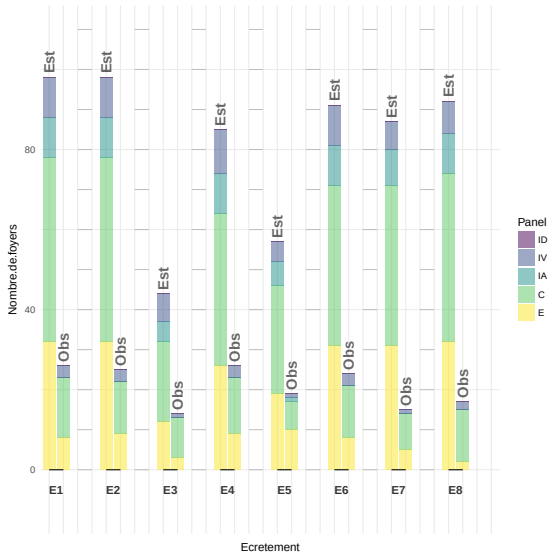

FIGURE 4.7 : Nombre de foyers ayant un contrat 6 kVA AC (Est) et ayant subi au moins une ouverture de breaker (Obs) selon le jour de l'écrêtement et pour chaque panel.

Panel	Écr. 1	Écr. 2	Écr. 3	Écr. 4	Écr. 5	Écr. 6	Écr. 7	Écr. 8
C	0.067	0.115	0.06	0.115	0	0	0.053	0.092
C	0.43	0.492	0.352	0.483	0.458	0.684	0.693	0.738
C	0.681	0.71	0.742	0.769	0.446	0.76	0.675	0.704
C	0.135	0.2	0.365	0.379	0.628	0.559	0.299	0.38
C	0.27	0.305	0.41	0.42	0.003	0.302	0.312	0.347
C	0.028	0.059	0.014	0.047	0.1	0.019	0.081	0.141
C	0.235	0.299	0.297	0.571	0.415	0.563	0.471	0.538
C	0.143	0.187	0.337	0.473	0.556	0.554	0.451	0.509
C	0.804	0.838	0.735	0.774	0.968	0.866	0.849	0.875
IV	0.21	0.281	0.231	0.388	0.637	0.141	0.11	0.161
IV	0.003	0.007	0.015	0.02	0.361	0.061	0.023	0.043
ID	0.243	0.283	0.138	0.255	0.752	0.465	0.368	0.41

TABLE 4.12 : Probabilités estimées de dépasser la puissance maximale en période d'écrêtement pour les foyers avec une puissance souscrite de 12 kVA pour l'hiver 2017/2018.

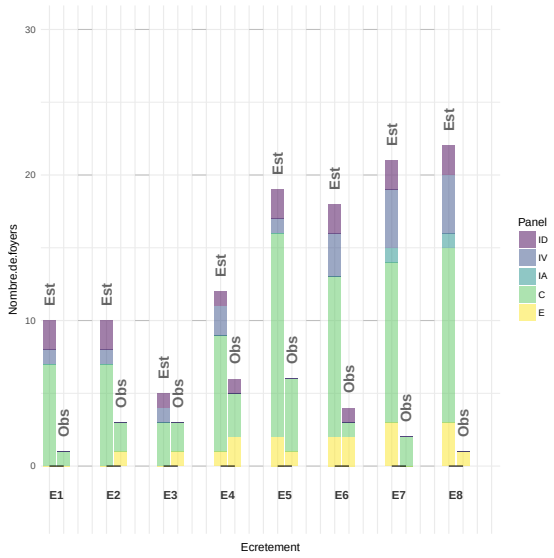

FIGURE 4.9 : Nombre de foyers ayant un contrat 9 kVA AC estimé à risque (Est) et ayant subi au moins une ouverture de breaker (Obs) selon le jour de l'écrêtement et pour chaque panel.

Récapitulatif

Dans cette section, nous allons faire un petit récapitulatif des résultats obtenus ainsi qu'une conclusion sur cette partie.

Nous donnons dans la Table 4.13 le nombre de foyers à risque pour chacun des écrêtements et pour chaque hiver selon le type de contrat ainsi que le nombre d'ouverture de breaker pour chaque écrêtement entre parenthèses. La probabilité estimée de dépasser la puissance maximale en période d'écrêtement correspond à une probabilité de dépassement si le foyer continue de consommer comme à son habitude. Or, nous considérons généralement plus de foyers à risque que nous n'observons d'ouverture de breaker. Nous pouvons faire l'hypothèse que les foyers ont changé leurs habitudes pour s'adapter aux écrêtements.

Il est facile de remarquer que plus les écrêtements sont durs, plus nous considérons de foyers à risque. Par exemple, si nous mettons en relation la Table 4.5 et la Figure 4.3 on peut voir que quand les foyers avec un contrat 9 kVA atteignent le talon, le nombre de foyers considérés à risque augmente considérablement.

Date d'écrêtement	3 kVA	6 kVA	9 kVA	12 kVA
11/1/2016	0 (0)	1 (3)	0 (0)	0 (0)
13/1/2016	0 (0)	8 (3)	7 (5)	0 (0)
18/1/2016	0 (0)	6 (2)	7 (2)	0 (0)
25/2/2016	0 (0)	9 (2)	8 (3)	0 (0)
1/3/2016	0 (0)	5 (0)	4 (2)	0 (0)
7/3/2016	0 (0)	15 (1)	17 (10)	3 (1)
18/3/2016	0 (0)	9 (8)	16 (2)	3 (1)
22/11/2016	1 (0)	72 (20)	34 (9)	2 (0)
14/12/2016	1 (0)	43 (24)	35 (12)	8 (1)
5/1/2017	1 (0)	36 (7)	34 (5)	5 (0)
17/1/2017	0 (0)	46 (15)	26 (22)	6 (1)
1/2/2017	0 (0)	30 (14)	28 (10)	4 (2)
28/2/2017	0 (0)	133 (32)	52 (21)	9 (1)
30/3/2017	1 (0)	48 (23)	47 (16)	8 (0)
9/11/2017	0 (0)	111 (31)	48 (18)	9 (3)
29/11/2017	0 (0)	111 (29)	55 (31)	10 (2)
19/12/2017	0 (0)	56 (16)	48 (17)	9 (5)
8/1/2018	1 (0)	101 (33)	63 (27)	10 (5)
2/2/2018	0 (0)	69 (22)	75 (25)	10 (2)
20/2/2018	0 (0)	105 (30)	72 (20)	9 (2)
13/3/2018	0 (0)	97 (28)	72 (5)	9 (1)
28/3/2018	0 (0)	103 (19)	74 (10)	11 (0)

TABLE 4.13 : Récapitulatif du nombre de foyers estimés à risque et ayant subi une ouverture de breaker (chiffre entre parenthèses).

Nombre de foyers	Hiver 2	Hiver 3	Différence en %
6 kVA à risque	58	94	62%
6 kVA ouverture de breaker	18	26	44%
9 kVA à risque	36	63	75%
9 kVA ouverture de breaker	15	19	27%

TABLE 4.14 : Moyenne du nombre de foyer estimés à risque et ayant subi une ouverture de breaker.

Si nous regardons le nombre moyen de foyers considérés à risque pour les écrêtements de l'hiver 2 et 3 présentés dans la Table 4.14, nous avons une augmentation du nombre de foyers considérés à risque plus importante que celle des foyers ayant subis une ouverture de breaker.

4.2.3 Proposition de réduction

Pour cette section, nous allons proposer une réduction de la puissance d'écrêtement de manière à ce que le moins d'expérimentateur possible soit considéré à risque. Pour ce faire, nous allons estimer la puissance correspondant à la probabilité qu'une mesure dépasse cette puissance pendant la période d'écrêtement.

À titre d'exemple, nous calculons la puissance minimale pour laquelle nous aurons 50% des foyers à risque. Il est possible de changer ce pourcentage pour une mise en situation de la méthode.

Hiver 1

Nous n'avons pas de données sur les foyers avec un contrat d'une puissance souscrite de 3 kVA pour l'hiver 2015/2016.

6 kVA

Nous allons utiliser les foyers avec une puissance souscrite de 6 kVA AC. Pour ceux-là, nous avons 54 foyers avec assez de données.

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	1335	2370	2319	2370
Pourcentage d'écrêtement associé (%)	77.75	60.5	61.35	60.5
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Puissance réduite (VA)	1325.5	2343	1907	
Pourcentage d'écrêtement associé (%)	77.91	60.95	68.22	

TABLE 4.15 : Puissance réduite médiane des foyers 6 kVA AC pour l'hiver 2015/2016.

Sur la Table 4.15, cette puissance médiane veut dire que sur le premier hiver et sur le premier écrêtement, nous avons 50% des foyers concernés qui ont une probabilité de dépasser 1335 VA inférieure à 0.08 et 50% des foyers concernés ont une probabilité supérieure à 0.08. Cela veut dire que si les foyers consomment à leurs habitudes, un écrêtement de 77.75% lors du 11 janvier 2016 de 17 à 19h aurait affecté la moitié des foyers 6 kVA AC.

9 kVA

La Table 4.16 présente la puissance estimée jusqu'à laquelle l'écrêtement est possible pour que 50% des foyers 9 kVA CE aient une probabilité de dépasser cette puissance inférieure à la probabilité d'avoir une mesure au-dessus de celle-ci.

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	3467	4438	4438	4438
Pourcentage d'écrêtement associé (%)	61.48	50.69	50.69	50.69
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Puissance réduite (VA)	3282	4438	3377	
Pourcentage d'écrêtement associé (%)	63.53	50.69	62.48	

TABLE 4.16 : Puissance réduite médiane des foyers 9 kVA CE pour l'hiver 2015/2016.

De la même façon, nous avons calculé dans la Table 4.17 la puissance d'écrêtement médiane associée à la probabilité d'avoir une mesure dépassant cette puissance pour les foyers avec une puissance souscrite de 9 kVA AC Nous lisons ces tableaux

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	2183	3100	3100	3149
Pourcentage d'écrêtement associé (%)	75.74	65.55	65.55	65.01
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Puissance réduite (VA)	2008	3100	2773	
Pourcentage d'écrêtement associé (%)	77.69	65.55	69.19	

TABLE 4.17 : Puissance réduite médiane des foyers 9 kVA AC pour l'hiver 2015/2016.

de la même manière que pour la Table 4.15. C'est-à-dire que si nous prenons le troisième écrêtement de l'hiver 2015/2016, nous avons estimé, pour les foyers 9 kVA CE, une puissance de 4438 VA. Cela veut dire que nous avons 50% des foyers 9 kVA CE qui ont une puissance estimée, associée à la probabilité 0.04, inférieure ou égale à 4438. En réalité, un écrêtement de 50% a été effectué, ce qui est quasiment égal au pourcentage d'écrêtement proposé à la Table 4.16. D'après notre estimation nous pouvons dire que si les foyers écrêtés ne changent pas leurs habitudes de consommation électrique, alors nous auront 50% des foyers dont la probabilité sera plus élevée que celle d'avoir une mesure dépassant la puissance écrêtée. Ces foyers seront donc considérés à risque. On retrouve quasiment le même nombre de foyers que celui Figure 4.3, la différence vient que l'écrêtement fût de 50% et non de 50.69%. Pour les foyers avec une puissance souscrite de 9 kVA AC, nous avons très peu de foyers concernés (10) sur l'hiver 2015/2016 mais nous avons quand même mis les résultats dans la Table 4.17. Nous pouvons remarquer ici que les puissances estimées pour les foyers AC sont inférieures à celles des foyers avec un mode de chauffage électrique.

Pour les foyers avec un contrat d'une puissance souscrite de 12 kVA, nous avons trop peu de foyers pour afficher le résultat des calculs de l'hiver 2015/2016.

Global

Sur la Table 4.18, nous avons calculé non plus une puissance estimée par type de contrat mais un pourcentage d'écrêtement global calculé à partir de la prise en compte de tous les types de contrats. Pour réaliser ceci, nous avons gardé le pourcentage de diminution de la puissance de chaque foyer et calculé la médiane sur l'ensemble d'entre eux. Si nous prenons comme exemple le deuxième écrêtement, le

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Pourcentage d'écrêtement global (%)	73.67	60.40	61.30	60.60
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Pourcentage d'écrêtement global (%)	74.00	60.60	67.46	

TABLE 4.18 : Pourcentage d'écrêtement médian estimé sur la totalité des foyers pour l'hiver 2015/2016.

pourcentage médian estimé est de 60.40%. Cela signifie que si les puissance écrêtées sont de 1.812, 3.624, 5.436 et 7.248 kVA pour, respectivement, les foyers avec une puissance souscrite de 3, 6, 9 et 12 kVA, alors nous auront considéré autant de foyers à risque que sans risque lors de cet écrêtement.

Hiver 2

Pour l'hiver 2016/2017, nous avons procédé de la même façon que pour l'hiver 2015/2016. Nous avons également trop peu de foyers avec un contrat d'une puissance souscrite de 3 kVA pour obtenir des résultats.

6 kVA

La Table 4.19 nous renseigne sur les puissances réduites médianes estimées pour les foyers avec une puissance souscrite de 6 kVA. On peut facilement remarquer que les puissances estimées des foyers CE sont supérieures à celles de ceux avec un autre mode de chauffage.

9 kVA

Nous pouvons voir sur la Table 4.20 les puissances médianes estimées pour les foyers avec une puissance souscrite de 9 kVA. De façon similaire, les foyers CE ont une puissance estimée supérieure à celle des foyers AC.

12 kVA

La Table 4.21 donne les résultats sur les foyers avec une puissance souscrite de 12 kVA dont nous disposons.

Foyers CE				
Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	2726	2194	2802	2405
Pourcentage d'écrêtement associé (%)	54.57	63.43	53.30	59.92
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Puissance réduite (VA)	1560	2603	1698	
Pourcentage d'écrêtement associé (%)	74.00	56.62	71.70	
Foyers AC				
Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	2431	2176	1752	1446
Pourcentage d'écrêtement associé (%)	59.48	63.73	70.80	75.90
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Puissance réduite (VA)	1146	1572	607	
Pourcentage d'écrêtement associé (%)	80.90	73.80	89.88	

TABLE 4.19 : Puissance réduite médiane des foyers 6 kVA pour l'hiver 2016/2017.

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	6683	5479	5857	4794
Pourcentage d'écrêtement associé (%)	44.31	54.34	51.19	60.05
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Puissance réduite (VA)	4681	5106	3774	
Pourcentage d'écrêtement associé (%)	60.99	57.45	68.55	

TABLE 4.21 : Puissance réduite médiane des foyers 12 kVA pour l'hiver 2016/2017.

Global

De la même façon que pour l'hiver 2015/2016, sur la Table 4.22, nous avons calculé un pourcentage d'écrêtement global sur tous les types de contrats. Pour réaliser ceci, nous avons gardé le pourcentage de diminution de chaque foyer et calculé la médiane sur l'ensemble d'entre eux.

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Pourcentage d'écrêtement global (%)	56.62	61.62	66.84	69.81
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Pourcentage d'écrêtement global (%)	75.65	68.50	85.69	

TABLE 4.22 : Pourcentage d'écrêtement médian estimé sur l'ensemble des foyers pour l'hiver 2016/2017.

Foyers CE				
Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	4960	4333	4696	3815
Pourcentage d'écrêtement associé (%)	44.88	51.86	47.82	57.61
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Puissance réduite (VA)	3354	3882	2964	
Pourcentage d'écrêtement associé (%)62.73	56.87	67.07		
Foyers AC				
Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	2884	2377	2229	1908
Pourcentage d'écrêtement associé (%)	67.96	73.59	75.23	78.80
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	
Puissance réduite (VA)	1248	1824	1014	
Pourcentage d'écrêtement associé (%)	86.13	79.73	88.73	

TABLE 4.20 : Puissance réduite médiane des foyers 9 kVA pour l'hiver 2016/2017.

Hiver 3

3 kVA

Pour l'hiver 2017/2018, nous avons peu de foyers (4) avec un contrat d'une puissance souscrite de 3 kVA mais nous avons tout de même affiché les résultats dans la Table 4.23.

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	1296	1296	1161	1296
Pourcentage d'écrêtement associé (%)	56.8	56.8	61.3	56.8
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	**Écr. 8**
Puissance réduite (VA)	1323	1296	1296	1296
Pourcentage d'écrêtement associé (%)	55.9	56.8	56.8	56.8

TABLE 4.23 : Puissance réduite médiane des foyers 3 kVA pour l'hiver 2017/2018.

6 kVA

Sur la Table 4.24, nous avons les pourcentages d'écrêtements estimés pour les foyers avec une puissance souscrite de 6 kVA sur l'hiver 2017/2018.

Foyers CE				
Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	2265	2265	1968	2265
Pourcentage d'écrêtement associé (%)	62.25	62.25	67.20	62.25
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	**Écr. 8**
Puissance réduite (VA)	1782	2265	2265	2265
Pourcentage d'écrêtement associé (%)	70.30	62.25	62.25	62.25
Foyers AC				
Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	1578	1578	1551	1578 8
Pourcentage d'écrêtement associé (%)	73.70	73.70	74.15	73.70
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	**Écr. 8**
Puissance réduite (VA)	1668	1578	1578	1578
Pourcentage d'écrêtement associé (%)	72.20	73.70	73.70	73.70

TABLE 4.24 : Puissance réduite médiane des foyers 6 kVA pour l'hiver 2017/2018.

On peut remarquer que les pourcentages d'écrêtements sont plus élevés pour les foyers AC que pour ceux CE.

9 kVA

Nous pouvons voir sur la Table 4.25 les puissances médianes estimées pour les foyers avec une puissance souscrite de 9 kVA.

Foyers CE				
Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	4003	4003	3738	4003
Pourcentage d'écrêtement associé (%)	55.52	55.52	58.47	55.52
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	**Écr. 8**
Puissance réduite (VA)	3827	4003	3924	3966
Pourcentage d'écrêtement associé (%)	57.48	55.52	56.40	55.93
Foyers AC				
Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	2028	2028	1985	2028
Pourcentage d'écrêtement associé (%)	77.47	77.47	77.94	77.47
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	**Écr. 8**
Puissance réduite (VA)	2091	2028	2028	2028
Pourcentage d'écrêtement associé (%)	76.77	77.47	77.47	77.47

TABLE 4.25 : Puissance réduite médiane des foyers 9 kVA pour l'hiver 2017/2018.

De façon similaire, les foyers CE ont une puissance estimée supérieure à celle des foyers AC.

12 kVA

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Puissance réduite (VA)	5454	5454	5190	5454
Pourcentage d'écrêtement associé (%)	54.55	54.55	56.75	54.55
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	**Écr. 8**
Puissance réduite (VA)	4292	5454	5454	5454
Pourcentage d'écrêtement associé (%)	64.23	54.55	54.55	54.55

TABLE 4.26 : Puissance réduite médiane des foyers 12 kVA pour l'hiver 2017/2018.

La Table 4.21 donne les résultats sur le peu de foyers avec une puissance souscrite de 12 kVA dont nous disposons.

Global

Pour finir, nous avons calculé le pourcentage d'écrêtement global sur tous les types de contrats. Pour réaliser ceci, nous avons gardé le pourcentage de diminution de chaque foyer et calculé la médiane sur l'ensemble d'entre eux.

Écrêtement	**Écr. 1**	**Écr. 2**	**Écr. 3**	**Écr. 4**
Pourcentage d'écrêtement global (%)	68.80	68.80	70.70	68.80
Écrêtement	**Écr. 5**	**Écr. 6**	**Écr. 7**	**Écr. 8**
Pourcentage d'écrêtement global (%)	70.12	68.80	68.80	68.80

TABLE 4.27 : Puissance réduite médiane globale pour l'hiver 2017/2018.

Commentaires

Nous remarquons que sur chacun des hivers, nous avons estimé une puissance d'écrêtement plus faible avec les foyers AC, cela nous réconforte quant au résultat de la classification effectuée sur le type de chauffage. De plus, nous remarquons qu'en général, les foyers CE ont un contrat en heure pleine/heure creuse, il n'est donc pas impensable de séparer le niveau d'écrêtement par type de contrat pour mettre en place ce modèle.

Nous proposons un moyen de calculer une puissance permettant de définir un pourcentage d'écrêtement en prenant en compte le nombre de foyers impactés.

Comme le temps de calcul est très rapide, ce modèle pourrait être utilisé pour définir le pourcentage d'écrêtement suivant un incident sur le réseau tout en prenant en compte le nombre de foyers qui seront considérés à risque. Par exemple, nous pourrons répondre à l'interrogation suivante : nous voulons que seulement 30% des foyers soient considérés à risque, quel est le pourcentage d'écrêtement associé ?

4.3 Étude des impacts des visites individuelles.

Cette section a pour but d'étudier l'impact éventuel des visites individuelles effectuées par ALOEN (actions de maitrise de la consommation électrique) sur la consommation électrique. Nous avons fait l'hypothèse de trois sortes d'impacts :

- court - une semaine
- moyen - un mois
- long - jusqu'à la prochaine visite.

Comme les visites n'ont pas eu lieu en même temps pour l'ensemble des foyers, nous ne pourrons pas tirer de conclusion sur le résultat que nous allons obtenir comme nous ne comparons pas les foyers dans les même conditions. En effet, un foyer peut partir en vacances la semaine suivant une visite et cela impactera nos résultats. Sur une durée plus longue, ces vacances auront un impact moins grand. Dans cette section, nous avons pris en compte un lien linéaire entre la température extérieure et la consommation électrique définie par le modèle Section 1.3. Il y a eu entre 2 et 4 visites pour 47 foyers répartis sur 2016-2017 :

1. entre le 12 janvier et 1er mars 2016
2. entre le 23 mars et 20 juin 2016
3. entre le 21 septembre et 7 décembre 2016
4. entre le 14 février et 2 mai 2017.

Il n'y a eu aucun foyer avec moins d'un mois d'écart entre deux visites. Pour illustrer de ce qui est fait dans cette section, nous avons pris l'exemple d'un foyer en regardant la semaine suivant sa visite.

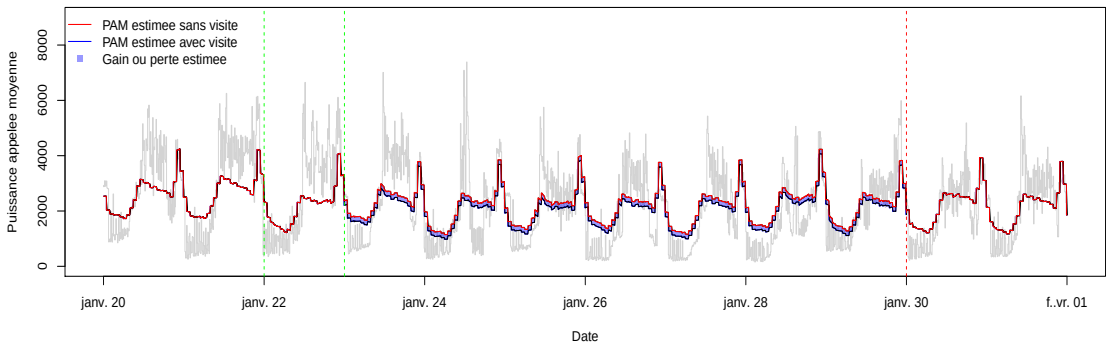

FIGURE 4.10 : Illustration de la puissance appelée estimée d'un foyer avec une puissance souscrite de 9 kVA et ayant une baisse significative de sa consommation durant la semaine suivant la visite. Le jour de la visite est le 22 janvier (pointillés verts) et la semaine suivante est du 23 au 30 janvier (entre pointillés verts et rouges).

Il est intéressant de regarder si la différence entre la PAM estimée sans visite (rouge) et celle avec visite (bleue) est significative. Pour qu'elle soit significative, il faut que le paramètre associé à la visite soit significativement différent de 0. Cette méthode a été généralisée pour le mois suivant suivant ainsi que pour la période entre deux visites.

Pour pouvoir regarder si la visite a eu un impact sur les consommations, nous avons effectué un test d'indépendance du χ^2. Ce test permet de comparer deux tableaux de contingence et de regarder si ils sont significativement différents l'un de l'autre.

Nous avons comparé ces tableaux avec celui des témoins. Nous savons que ceux-ci n'ont pas reçu de visites. Nous avons estimé la distribution de la variation de consommation des témoins en prenant en compte chaque semaine, puis le mois suivant la date moyenne des visites. Le tableau des foyers suivis individuellement a été ensuite comparé à cette distribution par un test d'indépendance du χ^2. La moyenne des pourcentages des foyers témoins pour chaque catégorie a été ajoutée pour chaque visite.

Nous avons cependant certaines visites dont le nombre de foyer exploitable est inférieur à 10. Le test du χ^2 n'a pas été effectué pour ces visites car ce nombre est trop faible.

Nous avons toujours des informations non exploitées dans les erreurs du modèle, c'est pourquoi nous considérons un modèle de chaîne de Markov à états cachés pour récolter un maximum d'information des erreurs. Il a été conseillé dans Durand et al. [32] d'utiliser une chaîne de Markov à états cachés à 7 états pour modéliser la dis-

tribution des erreurs. La distribution de chaque état étant choisie comme provenant d'une loi normale, chaque état est défini par une moyenne et un écart-type. Voici un exemple des états restitués pour un foyer pris au hasard.

TABLE 4.28 : Exemple de paramètres estimés de la distribution de chaque états pour un foyer.

Moyenne						
S_1	S_2	S_3	S_4	S_5	S_6	S_7
-0.88	-0.61	-0.32	0.09	0.34	0.52	1.08
écart-type						
S_1	S_2	S_3	S_4	S_5	S_6	S_7
0.22	0.11	0.12	0.19	0.12	0.20	0.37

Comparons l'histogramme des erreurs pour ce foyer avec la densité théorique de la loi de mixture gaussienne. La Figure 4.12 affiche la consommation du foyer

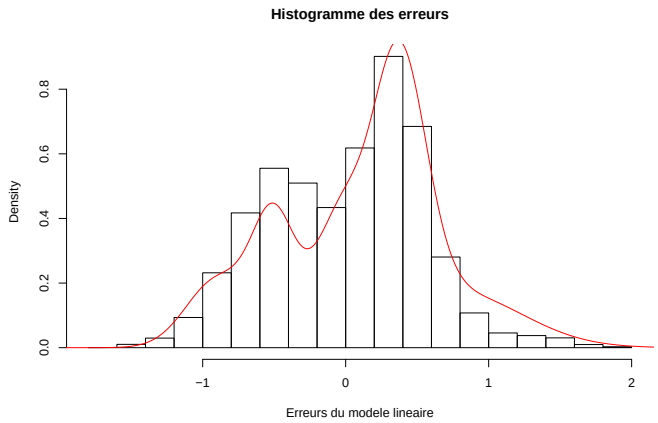

FIGURE 4.11 : Histogramme des erreurs et densité théorique de la distribution de mixture gaussienne pour un foyer 6 kVA avec un mode de chauffage principal non électrique.

pour quelques jours, les erreurs ainsi que les états avec la probabilité que les erreurs appartiennent à cet état la plus élevée. On remarque sur la figure 4.12 un appareil récurrent toutes les deux mesures, cela pourrait être le réfrigérateur. Lors de ces périodes (par exemple le mercredi matin), les erreurs sont alternativement positifs et négatifs, cela signifie que l'appareil fonctionne habituellement durant ces heures (pic positif suivi par un pic négatif dans les erreurs). De plus, nous remarquons un pic dans la consommation le jeudi et vendredi soir qui apparaît également dans les erreurs, cela signifie une consommation électrique différente de celle habituelle, et dans ce cas, plus élevée.

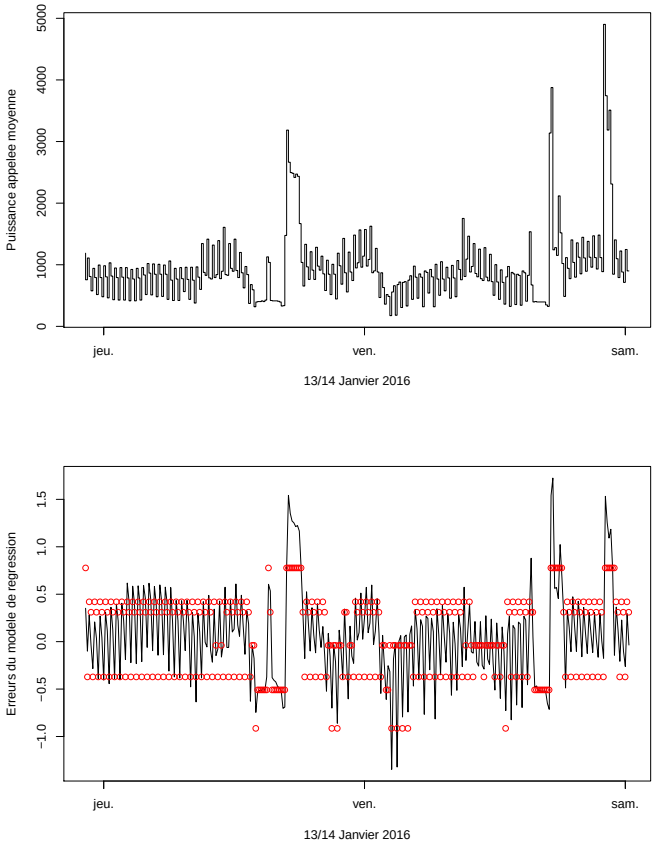

FIGURE 4.12 : Consommation électrique (haut) et erreurs du modèle (bas) avec la moyenne estimée des états d'appartenance en rouges pour un foyer 6 kVA avec un mode de chauffage principal non électrique.

La modélisation des erreurs par une chaîne de Markov à états cachés permet de voir des changements d'états entre les différentes heures, il nous manque cependant l'information des appareils mis en place pour liés les deux résultats. c'est pourquoi les résultats présents dans ce chapitre n'utilisent pas cette modélisation des erreurs.

4.3.1 Étude de la semaine suivant la visite individuelle.

Nous avons décidé de ne choisir que les foyers avec au moins 90% des mesures sur la période donnée. L'estimation de la consommation est plus fiable (si nous comparons celle estimée avec la consommation réelle) sur la période pour les foyers avec plus de 90% des mesures. L'écart entre l'estimé et le réel est trop important si nous gardons 80% des mesures ou moins.

Visite 1

Pour la première visite, nous n'avons que 17 foyers avec des données sur la semaine suivant la visite et seulement 10 d'entre eux ont plus de 90% des mesures sur la période étudiée.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	2	4	4
Pourc. Témoins	26.66%	45.52%	27.81%

TABLE 4.29 : Répartition des foyers selon la variation de consommation électrique pour la semaine suivant la première visite.

Le nombre de foyer n'est pas suffisant pour effectuer un test du χ^2.

Visite 2

Pour la deuxième visite, nous n'avons que 23 foyers avec des données sur la semaine suivant la visite et seulement 10 d'entre eux ont plus de 90% des mesures sur la période étudiée. Ces 10 foyers ne sont pas les mêmes que pour ceux présents dans la Table 4.29.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	1	3	5
Pourc. Témoins	26.46%	51.64%	21.90%

TABLE 4.30 : Répartition des foyers selon la variation de consommation électrique pour la semaine suivant la deuxième visite.

Comme pour le tableau concernant la première visite, nous avons ici trop peu de foyers pour effectuer un test du χ^2.

Visite 3

Pour la troisième visite, nous avons 42 foyers avec des données sur la semaine suivant la visite et 38 d'entre eux ont plus de 90% des mesures sur la période étudiée.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	18	14	6
Pourc. Témoins	32.02%	39.66%	28.32%

TABLE 4.31 : Répartition des foyers selon la variation de consommation électrique pour la semaine suivant la troisième visite.

Un test du χ^2 a été effectué sur cet échantillon en comparant avec la distribution des témoin et n'a pas conclut à un effet de la visite sur la consommation électrique.

Visite 4

Pour la dernière visite, nous avons 38 foyers avec des données sur la semaine suivant la visite et 34 d'entre eux ont plus de 90% des mesures sur la période étudiée.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	10	19	5
Pourc. Témoins	31.19%	43.25%	25.56%

TABLE 4.32 : Répartition des foyers selon la variation de consommation électrique pour la semaine suivant la quatrième visite.

Un test du χ^2 a été effectué entre les foyers suivis individuellement et la distribution des témoins et n'a pas conclut à un effet de la visite sur la consommation électrique.

Le test du χ^2 ne permettant pas de discriminer pour la semaine, nous allons maintenant regarder pour le mois suivant la visite. Les résultats peuvent être liés au fait que les visites sont étalées sur deux mois, nous pouvons comparer des semaines complètement différentes entre les foyers et une simple modification des habitudes d'un foyer durant la semaine peut avoir un impact important sur la consommation électrique (absence, invités, ...).

4.3.2 Étude du mois suivant la visite individuelle

Au lieu de regarder la semaine suivant la visite comme illustré sur la Figure 4.10, nous regardons le mois suivant la visite. Nous avons utilisé la même démarche que dans la Section 4.3.1. C'est-à-dire que nous regardons si le paramètre correspondant

à la visite est significativement différent de 0. Nous avons décidé de ne choisir que les foyers avec au moins 80% des mesures sur la période donnée. En effet, l'estimation de la consommation reste correcte (si nous comparons celle estimée avec la consommation réelle) sur le mois suivant la visite pour les foyers quand nous avons au moins de 80% des mesures. Comme nous estimons une période plus grande que dans la Section 4.3.1, nous pouvons utiliser un pourcentage moins élevé et avoir tout de même une bonne estimation de la consommation.

Visite 1

Pour la première visite, nous avons 27 foyers avec des données sur le mois suivant la visite et seulement 9 d'entre eux ont plus de 80% des mesures sur la période étudiée.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	6	1	2
Pourc. Témoins	27.18%	35.92%	36.89%

TABLE 4.33 : Répartition des foyers selon la variation de consommation électrique pour le mois suivant la première visite.

Comme nous pouvons le voir avec la Table 4.33, nous avons très peu de foyers possédant des données sur le mois suivant la première visite. Nous n'avons pas effectué de test du χ^2 pour cette visite.

Visite 2

Pour la deuxième visite, nous avons 35 foyers avec des données sur le mois suivant la visite et 10 d'entre eux ont plus de 80% des mesures sur la période étudiée.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	2	5	3
Pourc. Témoins	36.76%	44.12%	19.12%

TABLE 4.34 : Répartition des foyers selon la variation de consommation électrique pour le mois suivant la deuxième visite.

Nous n'avons pas effectué de test du χ^2 sur cette visite dû au faible nombre de données.

Visite 3

Pour la troisième visite, nous avons 44 foyers avec des données sur le mois suivant la visite et 40 d'entre eux ont plus de 80% des mesures sur la période étudiée.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	12	17	11
Pourc. Témoins	6.76%	90.70%	2.54%

TABLE 4.35 : Répartition des foyers selon la variation de consommation électrique pour le mois suivant la troisième visite.

Le test du χ^2 a été effectué sur les données du mois suivant la troisième visite. Le test conclut à une différence entre la distribution des foyers suivis individuellement et les témoins. En revanche, si nous regardons la Table 4.35, la variation de la consommation électrique pour le mois suivant la troisième visite semble distribué uniformément. Un test du χ^2 entre la répartition des foyers suivis individuellement et l'uniformité de la distribution permet de confirmer cette hypothèse. Nous ne pouvons pas rejeter l'uniformité de la distribution des foyers individuels selon leurs variation de consommation.

Visite 4

Pour la quatrième visite, nous avons 38 foyers avec des données sur le mois suivant la visite et 37 d'entre eux ont plus de 80% des mesures sur la période étudiée.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	17	15	5
Pourc. Témoins	6.93%	88.80%	4.27%

TABLE 4.36 : Répartition des foyers selon la variation de consommation électrique pour le mois suivant la dernière visite.

Un test du χ^2 a été effectué et a conclut à une différence entre la distribution des foyers suivis individuellement et des témoins (p-value < 0.05). De plus le test du χ^2 a également conclut à une différence avec la distribution uniforme. On peut conclure à un effet de la quatrième visite sur la variation de la consommation sur le mois suivant cette visite.

Le test du χ^2 n'a pas permis de discriminer sur les données du mois suivant la troisième visite mais il a permis de détecter une répartition différente sur la quatrième visite. Nous allons maintenant regarder si nous pouvons détecter un effet pour la variation de la consommation entre deux visites.

4.3.3 Étude entre deux visites

Comme la durée entre deux visites peut varier selon les foyers, il a été décidé de garder les foyers où l'estimation de la consommation électrique était proche de celle observée. Pour réaliser ceci, nous avons regardé l'erreur relative absolue (ARE) entre les deux consommations (i.e. ARE) et pris les foyers où l'erreur était inférieure à 20%. Cette valeur a été choisie arbitrairement pour éviter d'avoir trop ou trop peu de foyers.

Visite 1

Pour la première visite, nous avons 33 foyers avec des données entre la première et la deuxième visite et nous avons gardé 7 d'entre eux.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	5	1	1
Pourc. Témoins	39.22%	27.45%	33.33%

TABLE 4.37 : Répartition des foyers selon la variation de consommation électrique entre la première et la deuxième visite.

Nous avons ici trop peu de foyers pour effectuer un test du χ^2.

Visite 2

Pour la deuxième visite, nous avons 42 foyers avec des données entre la deuxième et la troisième visite et nous avons gardé 29 d'entre eux.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	7	15	7
Pourc. Témoins	26.87%	47.14%	25.99%

TABLE 4.38 : Répartition des foyers selon la variation de consommation électrique entre la deuxième et la troisième visite.

Le test du χ^2 n'a pas conclut à un effet de la deuxième visite sur la consommation électrique entre la deuxième et la troisième visite.

Visite 3

Pour la troisième visite, nous avons 44 foyers avec des données entre la troisième et la quatrième visite et nous avons gardé 36 d'entre eux.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	6	16	14
Pourc. Témoins	19.61%	45.42%	34.97%

TABLE 4.39 : Répartition des foyers selon la variation de consommation électrique entre la troisième et la quatrième visite.

Le test du χ^2 n'a pas conclut à un effet de la visite sur la consommation électrique entre la troisième et la quatrième visite.

Visite 4

Pour la quatrième visite, nous avons 36 foyers avec des données après la quatrième et nous avons gardé 31 d'entre eux.

	Baisse significative	Pas de baisse ni augmentation significative	Augmentation significative
Nb Foyers	8	18	5
Pourc. Témoins	36.81%	40.07%	23.13%

TABLE 4.40 : Répartition des foyers selon la variation de consommation électrique après la quatrième visite.

Le test du χ^2 n'a pas conclut à un effet de la visite sur la consommation électrique après la dernière visite.

4.3.4 Commentaires

Nous avons effectué un test d'indépendance du χ^2 pour comparer la distribution des foyers suivis individuellement avec les témoins pour trois périodes différentes. Nous avons pour quasiment toutes les visites la conclusion qu'il n'y a pas de différence de distribution entre les deux panels. Nous avons cependant détecté une différence de distribution pour le mois suivant la quatrième visite.

Les résultats sont mis en question dû au faible nombre de foyers, un événement atypique peu influencer les résultats sur un échantillon de cette taille.

Nous avons regardé l'impact sur la semaine ou le mois suivant chaque visite. mais nous n'avons pas regardé un impact longitudinal des visites sur la consommation électrique. Il nous manque les données antérieures à 2016 pour pouvoir suivre l'évolution de la consommation sur un périmètre plus grand.

Conclusion et perspectives

L'objectif de cette thèse était de proposer un modèle permettant d'estimer les données de courbe de charge dans le but de répondre à plusieurs problématiques. Cet objectif s'inscrit à la suite du développement des compteurs électriques nouvelle génération "Linky". Un projet appelé SOLENN suit notamment cette pose et a cherché à proposer des solutions de maîtrise de la consommation électrique pour les particuliers.

Pour faciliter la lecture du document, nous rappelons dans le chapitre 1 les différents concepts utilisés dans celui-ci. Un rappel sur la théorie des valeurs extrêmes est tout d'abord donné puis un rappel sur l'analyse de survie et le modèle de Cox est proposé. Un modèle utilisé pour répondre à une des problématiques du projet est également présenté ainsi que les données sur lesquelles le travail a été effectué.

Lors du Chapitre 2, nous proposons un article qui fait suite à la mise en place et la publication d'un package R nommé **extremefit**. Cet article présente l'utilisation du package qui consiste à estimer les probabilités d'événements rares et des quantiles extrêmes conditionnels. Il s'agit d'un travail suivant la thèse de Pham [3].

Le Chapitre 3 introduit un modèle basé sur le modèle de Cox ainsi que sur la théorie des valeurs extrêmes. Il permet d'introduire des covariables dans un modèle à valeurs extrêmes. Les estimateurs des paramètres du modèle sont proposés ainsi que la convergence du paramètre de la distribution de Pareto. Plusieurs simulations ont été effectuées pour étudier le comportement des différents estimateurs.

Le Chapitre 4 présente une partie des résultats de l'application des modèles sur les données de courbes de charge. Deux problématiques du projet SOLENN sont abordées lors de ce chapitre. La première problématique fut de vérifier si les méthodes de maîtrise de la consommation électrique mises en place ont un impact sur la consommation des particuliers. Nous n'avons pas été en moyen de répondre à cette question dû au faible nombre de données dont nous disposions. La deuxième problématique fut de proposer un modèle permettant d'estimer si un foyer est à risque lors de l'écrêtement ciblé. Nous avons proposé un moyen de calculer un pourcentage d'écrêtement en fonction du nombre de foyers considérés à risque.

À la suite de cette thèse, il serait intéressant de proposer un modèle des quantiles de régression en prenant en compte les valeurs extrêmes. Le modèle de Cox a ses limites lorsque les fonctions de hasards ne sont pas proportionnelles en fonction des covariables. Adapter le modèle des quantiles de régression permettrait de ne plus se préoccuper de cette hypothèse.

Le travail à suivre est l'inclusion du modèle proposé dans le chapitre 3 dans le package **extremefit**.

Bibliographie

[1] N. Allab. *Détection d'anomalies et de ruptures dans les séries temporelles. Applications à la gestion de production de l'électricité.* PhD thesis, Paris 6, 2016.

[2] M. Sanquer. *Détection et caractérisation de signaux transitoires : application à la surveillance de courbes de charge.* PhD thesis, Grenoble, 2013.

[3] Q.-K. Pham. *Estimation non paramétrique adaptative dans la théorie des valeurs extrêmes : application en environnement.* PhD thesis, Université de Bretagne Sud, 2015.

[4] P. Ndao. *Modélisation de valeurs extrêmes conditionnelles en présence de censure.* PhD thesis, thése de doctorat, université Gaston berger de Saint'Louis, 2015.

[5] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of Extremes : Theory and Applications.* John Wiley & Sons, 2004.

[6] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2) :187–220, 1972. ISSN 00359246. URL http://www.jstor.org/stable/2985181.

[7] W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4) :945–966, 1972.

[8] O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

[9] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282) :457–481, 1958.

[10] T. A. Buishand, L. de Haan, and C. Zhou. On spatial extremes : with application to a rainfall problem. *The Annals of Applied Statistics*, 2(2) :624–642, 2008.

[11] L. Gardes and S. Girard. Conditional extremes from heavy-tailed distributions : an application to the estimation of extreme rainfall return levels. *Extremes*, 13 : 177–204, 2010.

[12] J. Danielsson and C. G. de Vries. Tail index and quantile estimation with very high frequency data. *Journal of empirical Finance*, 4(2) :241–257, 1997.

[13] A. J. McNeil. Calculating quantile risk measures for financial return series using extreme value theory. *ETH Zürich, Departement Mathematik*, 1998.

[14] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events : for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

[15] R. Gencay and F. Selcuk. Extreme value theory and value-at-risk : relative performance in emerging markets. *International Journal of Forecasting*, 20(2) : 287–303, 2004.

[16] A. J. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN bulletin*, 27(01) :117–137, 1997.

[17] H. Rootzén and N. Tajvidi. Extreme value statistics and wind storm losses : a case study. *Scandinavian Actuarial Journal*, 1997(1) :70–94, 1997.

[18] J. Worms and R. Worms. New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. *Extremes*, 17(2) :337–358, 2014.

[19] J. Worms and R. Worms. Extreme value statistics for censored data with heavy tails under competing risks. *Metrika*, 81(7) :849–889, 2018.

[20] J. Einmahl, A. Fils-Villetard, A. Guillou, et al. Statistics of extremes under random censoring. *Bernoulli*, 14(1) :207–227, 2008.

[21] G. Stupfler. Estimating the conditional extreme-value index under random right-censoring. *Journal of Multivariate Analysis*, 144 :1–24, 2016.

[22] J. El Methni, L. Gardes, S. Girard, et al. Kernel estimation of extreme regression risk measures. *Electronic journal of statistics*, 12(1) :359–398, 2018.

[23] A. Daouia, S. Girard, G. Stupfler, et al. Extreme m-quantiles as risk measures : From l1 to lp optimization. *Bernoulli*, 25(1) :264–309, 2019.

[24] J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[25] J. P. Klein and M. L. Moeschberger. *Survival analysis : techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[26] T. M. Therneau and P. M. Grambsch. *Modeling survival data : extending the Cox model*. Springer Science & Business Media, 2013.

[27] B. W. Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American statistical association*, 69(345) : 169–173, 1974.

[28] P. K Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. Statistical models based on counting processes. pages 176–331. Springer, 1993.

[29] N. E. Breslow. Discussion of professor Cox's paper. *J Royal Stat Soc B*, 34 : 216–217, 1972.

[30] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

[31] J.-J. Droesbeke and G. Saporta. *Approches non paramétriques en régression*. Editions Technip, 2011.

[32] J.-B. Durand, L. Bozzi, G. Celeux, and C. Derquenne. *Analyse de courbes de consommation électrique par chaînes de Markov cachées*. PhD thesis, INRIA, 2003.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[34] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2) :195–239, 1984.

[35] Z. Ghahramani and M. I. Jordan. Learning from incomplete data. 1995.

[36] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2) :181–214, 1994.

[37] C. Bishop, C. M. Bishop, et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[38] C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[39] J. A. Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510) :126, 1998.

[40] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7) :719–725, 2000.

[41] G. Durrieu, I. Grama, Q. Pham, Jaunâtre K., and J.-M. Tricot. extremefit : A package for extreme quantiles. *Journal of Statistical Software*, to appear, 2017.

[42] P. Sharma, M. Khare, and S. P. Chakrabarti. Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. *Transportation Research Part D : Transport and Environment*, 4(3) : 201–216, 1999.

[43] A. Davison and R.L. Smith. Models for exceedances over high thresholds. *J. Roy. Statist. Soc. B*, 52(3) :393–442, 1990.

[44] R. W. Katz, M. B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in water resources*, 25(8) :1287–1304, 2002.

[45] P. Embrechts, S. I. Resnick, and G. Samorodnitsky. Extreme value theory as a risk management tool. *North American Actuarial Journal*, 3(2) :30–41, 1999.

[46] D. Sornette, L. Knopoff, Y. Kagan, and C. Vanneste. Rank-ordering statistics of extreme events : application to the distribution of large earthquakes. *Journal of Geophysical Research : Solid Earth*, 101(B6) :13883–13893, 1996.

[47] P. Hall and A. H. Welsh. Adaptive estimates of parameters of regular variation. *Ann. Statist.*, 13(1) :331–341, 1985.

[48] H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Process. Appl.*, 75 :149–172, 1998.

[49] A. Guillou and P. Hall. A diagnostic for selecting the threshold in extreme-value analysis. *J. Roy. Statist. Soc. Ser. B*, 63 :293–305, 2001.

[50] R. Huisman, C. G. Koedijk, C. J. M. Kool, and F. Palm. Tail index estimates in small samples. *Journal of Business and Economic Statistics*, 19(2) :208–216, 2001.

[51] I. Grama and V. Spokoiny. Statistics of extremes by oracle estimation. *Annals of Statistics*, 36(4) :1619–1648, 2008.

[52] I. Grama and V. Spokoiny. Pareto approximation of the tail by local exponential modeling. *Bulletin of Academi of Sciences of Moldova*, 53(1) :1–22, 2007.

[53] J. El Methni, L. Gardes, S. Girard, and A. Guillou. Estimation of extreme quantiles from heavy and light tailed distributions. *Journal of Statistical Planning and Inference*, 142(10) :2735–2747, 2012.

[54] G. Durrieu, I. Grama, Q. Pham, and J.-M. Tricot. Nonparametric adaptive estimator of extreme conditional tail probabilities quantiles. *Extremes*, 18 : 437–478, 2015.

[55] J.G. Staniswalis. The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84(405) :276–283, 1989.

[56] C. Loader. *Local regression and likelihood*. Springer-Verlag, 2006.

[57] R Development Core Team. *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL http://www.R-project.org.

[58] E. Gilleland, M. Ribatet, and A.G. Stephenson. A software review for extreme value analysis. *Extremes*, 16(1) :103–119, 2013.

[59] M. Ribatet and C. Dutang. *POT : Generalized Pareto Distribution and Peaks Over Threshold*, 2016. URL http://CRAN.R-project.org/package=POT. R package version 1.1-6.

[60] M. Ribatet, R. Singleton, and R Core team. ***SpatialExtremes*** *: Modelling Spatial Extremes*, 2011. URL http://spatialextremes.r-forge.r-project.org/. R package version 2.0-2.

[61] D. Wuertz and many others. ***fExtremes*** *: Rmetrics - Modelling Extreme Events in Finance*, 2013. URL http://CRAN.R-project.org/package=fExtremes. R package version 3010.81.

[62] M. Hofert, I. Kojadinovic, M. Maechler, and J. Yan. ***copula*** *: Multivariate Dependence with Copulas*, 2010. URL http://copula.r-forge.r-project.org/. R package version 0.999-14.

[63] A. Stephenson and C. Ferro. ***evd*** *: Functions for extreme value distributions*, 2002. URL http://CRAN.R-project.org/package=evir. R package version 2.3-0.

[64] A. Stephenson and M. Ribatet. ***evdbayes*** *: Bayesian Analysis in Extreme Value Theory*, 2010. URL http://CRAN.R-project.org/package=evdbayes. R package version 1.1-1.

[65] E. Gilleland. ***extRemes*** *: Extreme Value Analysis*, 2011. URL http://CRAN.R-project.org/package=extRemes. R package version 2.0-5.

[66] B. Pfaff, A. McNeil, and A. Stephenson. ***evir*** *: Extreme Values in R*, 2008. URL http://CRAN.R-project.org/package=evir. R package version 1.7-3.

[67] J. R. M. Hosking. ***lmom*** *: L-moments*, 2009. URL http://CRAN.R-project.org/package=lmom. R package version 2.5.

[68] H. Southworth and J. E. Heffernan. ***texmex*** *: Statistical modelling of extreme values*, 2010. URL http://code.google.com/p/texmex/. R package version 2.1.

[69] E. Simiu and N. A. Heckert. Extreme wind distribution tails : a peaks over threshold approach. *Journal of Structural Engineering*, 122(5) :539–547, 1996.

[70] D. Tran, P. Ciret, A. Ciutat, G Durrieu, and J.-C. Massabuau. Estimation of potential and limits of bivalve closure response to detect contaminants : application to cadmium. *Environmental Toxicology and Chemistry*, 22(4) :914–920, 2003.

[71] C. Chambon, A. Legeay, G. Durrieu, P. Gonzalez, P. Ciret, and J.-C. Massabuau. Influence of the parasite worm polydora sp. on the behaviour of the oyster crassostrea gigas : a study of the respiratory impact and associated oxidative stress. *Marine Biology*, 152(2) :329–338, 2007.

[72] M. Sow, G. Durrieu, and L. Briollais. Water quality assessment by means of HFNI valvometry and high-frequency data modeling. *Environmental Monitoring and Assessment*, 182(1-4) :155–170, 2011.

[73] F. G. Schmitt, M. De Rosa, G. Durrieu, M. Sow, P. Ciret, D. Tran, and J. C. Massabuau. Statistical study of bivalve high frequency microclosing behavior : scaling properties and shot noise modeling. *International Journal of Bifurcation and Chaos*, 21(12) :3565–3576, 2011.

[74] L. J. Jou and C. M. Liao. A dynamic artificial clam (corbicula fluminea) allows parcimony on-line measurement of waterborne metals. *Environmental Pollution*, 144 :172–183, 2006.

[75] R. Coudret, G. Durrieu, and J. Saracco. Comparison of kernel density estimators with assumption on number of modes. *Communication in Statistics - Simulation and Computation*, 44(1) :196–216, 2015.

[76] R. Azaïs, G. Coudret, and G. Durrieu. A hidden renewal model for monitoring aquatic systems biosensors. *Environmetrics*, 25 :189–199, 2014.

[77] G. Durrieu, Q.-K. Pham, A.-S. Foltête, V. Maxime, I. Grama, V. Le Tilly, H. Duval, J.-M. Tricot, C. B. Naceur, and O. Sire. Dynamic extreme values modeling and monitoring by means of sea shores water quality biomarkers and valvometry. *Environmental Monitoring and Assessment*, 188(7) :1–8, 2016.

[78] V. Bianco, O. Manca, and S. Nardini. Electricity consumption forecasting in italy using linear regression models. *Energy*, 34(9) :1413–1421, 2009.

[79] M. Ranjan and V. K. Jain. Modelling of electrical energy consumption in delhi. *Energy*, 24(4) :351–361, 1999.

[80] J. L. Harris and L.-M. Liu. Dynamic structural analysis and forecasting of residential electricity consumption. *International Journal of Forecasting*, 9(4) : 437–455, 1993.

[81] S. Bercu and F. Proïa. A sarimax coupled modelling applied to individual load curves intraday forecasting. *Journal of Applied Statistics*, 40(6) :1333–1348, 2013.

[82] P. K Andersen, O. Borgan, R. D Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[83] J. Crowley and M. Hu. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72(357) :27–36, 1977.

[84] E. W. Lee, L. J. Wei, D. A. Amato, and S. Leurgans. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival analysis : state of the art*, pages 237–247. Springer, 1992.

[85] J. Beirlant, A. Guillou, G. Dierckx, and A. Fils-Villetard. Estimation of the extreme value index and extreme quantiles under random censoring. *Extremes*, 10(3) :151–174, 2007.

[86] J. Beirlant, A. Guillou, and G. Toulemonde. Peaks-over-threshold modeling under random censoring. *Communications in Statistics ?Theory and Methods*, 39(7) :1158–1179, 2010.

[87] P. Ndao, A. Diop, and J.-F. Dupuy. Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. *Computational Statistics & Data Analysis*, 79 :63–79, 2014.

[88] P. Ndao, A. Diop, and J.-F. Dupuy. Nonparametric estimation of the conditional extreme-value index with random covariates and censoring. *Journal of Statistical Planning and Inference*, 168 :20–37, 2016.

[89] I. Grama, J.-M. Tricot, and J.-F. Petiot. Estimation of the extreme survival probabilities from censored data. *BULETINUL ACADEMIEI DE STIINTE A REPUBLICII MOLDOVA. MATEMATICA*, 74(1) :33–62, 2014.

[90] B. M. Hill et al. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 3(5) :1163–1174, 1975.

[91] P. Hall. On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 37–42, 1982.

[92] P. Hall and A. H. Welsh. Best attainable rates of convergence for estimates of parameters of regular variation. *The Annals of Statistics*, pages 1079–1084, 1984.

[93] H. Drees. Optimal rates of convergence for estimates of the extreme value index. *Annals of Statistics*, pages 434–448, 1998.

[94] D. P. Byar. The veterans administration study of chemoprophylaxis for recurrent stage i bladder tumours : comparisons of placebo, pyridoxine and topical thiotepa. In *Bladder tumors and other topics in urological oncology*, pages 363–370. Springer, 1980.

[95] L.-J. Wei, D. Y. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, 84(408) :1065–1073, 1989.

[96] D. R. Cox and E. J. Snell. Analysis of binary data. 2nd. *London : Chapman and HalI*, 1989.

[97] P. Grambsch and T. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3) :515–526, 1994.