AVERTISSEMENT

LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n°92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992
http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

Université Toulouse 1 Capitole

Toulouse School of Economics

École Doctorale Mathématiques,

Informatique et Télécommunications de Toulouse

MÉMOIRE EN VUE DE L'OBTENTION DE

L'HABILITATION À DIRIGER DES RECHERCHES

SPÉCIALITÉ : STATISTIQUE

CONTRIBUTIONS À L'ANALYSE STATISTIQUE DES DONNÉES CENSURÉES À DROITE

présenté et soutenu le 2 février 2018 par

Eve Leconte

devant le jury composé de

Laurent Bordes	Université de Pau et des Pays de l'Adour	Rapporteur
Agathe Guilloux	Université d'Evry Val d'Essonne	Rapportrice
Ingrid Van Keilegom	Université Catholique de Louvain	Rapportrice
Christine Thomas-Agnan	Université Toulouse 1 Capitole	Professeur référent
Jean-Yves Dauxois	INSA – Université de Toulouse	Examinateur
Thomas Filleron	Institut Claudius Regaud	Examinateur
Karen Leffondré	Université de Bordeaux	Examinatrice

A Bertrand,

A Roman, qui fait ses premiers pas
dans le monde de la recherche,

A Galaad, qui cherche à tout faire,

A Eléa, qui cherche encore sa voie.

Remerciements

En premier lieu, je tiens à exprimer ma gratitude à Christine Thomas-Agnan pour m'avoir accompagnée et encouragée durant toutes ces années et plus particulièrement ces derniers mois dans les diverses étapes de cette habilitation, depuis la rédaction du mémoire jusqu'à la préparation de la soutenance.

Je remercie chaleureusement Laurent Bordes, Agathe Guilloux et Ingrid Van Keilegom d'avoir accepté d'être les rapporteurs de ce travail et d'avoir pris le temps d'évaluer ce document de synthèse. Merci pour leurs remarques et conseils. Je remercie également Jean-Yves Dauxois, Thomas Filleron et Karen Leffondré d'avoir accepté de faire partie de mon jury de soutenance.

La recherche est un travail d'équipe et je remercie tous mes collaborateurs. Que ceux qui ne sont pas cités nommément ci-après ne m'en tiennent pas rigueur.

Merci d'abord à Sandrine Casanova, ma collaboratrice de prédilection et mon amie de vingt ans avec laquelle nous partageons tout, ou presque, y compris les chambres d'hôtels. Merci Sandrine pour ton soutien, ta disponibilité et ta bonne humeur.

Merci à Andrew Kramar de m'avoir mise en relation avec Thomas Filleron. Notre collaboration était motivée à la base par un projet commun : passer tous les deux notre habilitation à diriger des recherches à plus ou moins court terme. C'est bientôt chose faite, Thomas m'ayant devancée de trois ans. Merci Thomas pour m'avoir associée à l'encadrement de tes stagiaires et de tes étudiants en thèse. C'est un réel plaisir de collaborer avec toi et de pouvoir appliquer la méthodologie des données censurées au domaine médical qui m'a toujours intéressée.

Je remercie vivement Anne Ruiz-Gazen de m'avoir initiée à la recherche en sondages et d'être toujours disponible pour répondre à mes questions. Merci Anne pour ton soutien et ton amitié.

Merci à Yves Aragon, dont les jeux de mots me manquent depuis son départ à la retraite. Merci à Gérard Derzko qui a eu la patience de m'expliquer plusieurs fois ce que je ne comprenais pas toujours du premier coup. Merci à Jean-François Dupuy pour notre collaboration courte mais fructueuse. Merci à Pascal Maussion de nous avoir associées Anne et moi aux travaux de thèse de Farah. Enfin, j'ai une pensée particulière pour Christophe Cazaux qui nous a quittés bien trop tôt.

Je remercie tout particulièrement Philippe Besse pour m'avoir associée aux travaux de son étudiante en thèse Marie Walschaerts : outre le fait de m'avoir fait découvrir le domaine de la sélection de variables, il a aussi contribué à la création d'une amitié durable avec Marie. Marie, j'ai beaucoup apprécié notre collaboration j'espère qu'elle va pouvoir continuer malgré ton emploi du temps très chargé.

Merci aussi aux étudiants avec lesquels j'ai pris beaucoup de plaisir à travailler : Ibtissem, Agnès, Charlène, Serge, Soufiane, Bastien, Farah et maintenant Julia.

Pendant vingt ans, j'ai partagé mon bureau avec Isabelle Dubec, toujours d'humeur égale. Ce fut vingt ans d'entente parfaite sans la moindre divergence. Merci Isabelle pour ta présence sereine à mes côtés durant toutes ces années.

Merci à Thibaut, toujours prêt à m'aider pour résoudre mes problèmes de programmation avec le logiciel `R`. Merci aussi aux secrétaires de TSE, en particulier Aline, pour sa gentillesse et son aide logistique au quotidien.

Ce mémoire n'aurait pas eu le même aspect [1] sans l'aide de Nathalie Villa-Vialaneix, la fée de l'informatique — entre autres qualités — que je remercie vivement pour m'avoir

1. Concernant les éventuelles fautes de frappes résiduelles de ce mémoire — et l'expérience montre qu'il en reste toujours —, merci de vous adresser à mon amie Amandine, que je remercie pour m'avoir gentiment proposé — mais trop tard malheureusement — une relecture attentive de ce document ;-).

fourni le code source LaTeX de son mémoire d'HDR et permis ce plagiat formel. En amont, je remercie toutes les personnes qui développent ces outils précieux et les mettent librement à la disposition des chercheurs.

Pour finir, je voudrais remercier mon mari Bertrand, présent depuis le début de mes années de recherche, pour son amour et son soutien sans faille. Merci aussi à Eléa, ma grande fille de 13 ans, pour son autonomie qui m'a permis de dégager du temps pour mener à bien ce travail de synthèse. Enfin, merci à Callista, dont la zénitude féline m'invite au quotidien à relativiser mes problèmes d'humaine.

Table des matières

Présentation générale des travaux

Ce mémoire est la synthèse des travaux de recherche que j'ai menés ces vingt dernières années, essentiellement dans le cadre de l'analyse des données censurées à droite. La liste de mes publications numérotées se trouve à l'annexe B et j'utiliserai leur numéro pour y faire référence.

La plupart de mes travaux gravitent autour de trois axes de recherche, à savoir :

1. l'estimation non paramétrique de fonctions de répartition et de leurs fonctionnelles,
2. l'analyse de données censurées multivariées : événements répétés et/ou risques concurrents,
3. la sélection de variables dans les modèles de survie en grande dimension.

Outre l'analyse des données censurées, qui est le fil rouge commun à la quasi-totalité de mes travaux, les développements que je vais décrire ont fait appel à des techniques et des résultats provenant d'autres domaines de la statistique : estimation non paramétrique, théorie des sondages, sélection de modèles... Ils ont été rendus possibles grâce à des collaborations avec des spécialistes de ces domaines qui ont fait appel à moi pour mes compétences dans le domaine des données censurées et m'ont associée à l'encadrement et aux travaux de leurs étudiants en thèse : Christine Thomas-Agnan, enseignant-chercheur à UT1 Capitole, pour la thèse de Sandrine Casanova, Philippe Besse, enseignant-chercheur à l'INSA de Toulouse, pour la thèse de Marie Walschaerts, Thomas Filleron, méthodologiste-biostatisticien à l'Institut Claudius Regaud, pour la thèse de Serge Somda.

Le contenu des trois axes thématiques de recherche est résumé ci-dessous.

1. Estimation non paramétrique de fonctions de répartition (fdr) et de leurs fonctionnelles

Estimation de la fdr et de ses quantiles

Avec Sandrine Casanova, alors PRAG et étudiante en thèse sous la direction de Christine Thomas-Agnan, nous avons proposé dans [3] des estimateurs non paramétriques doublement lisses de la fdr conditionnelle et des quantiles d'une variable censurée à droite, c'est-à-dire lisses par rapport à la variable d'intérêt et par rapport à la cova-

riable. Nous montrons que ce double lissage permet d'améliorer les performances des estimateurs.

Estimation d'ordres-expectiles

Dans [6], nous proposons des estimateurs non paramétriques des ordres-expectiles conditionnels et montrons leur intérêt pour obtenir un classement des individus ajusté sur leurs caractéristiques. Nous appliquons ces estimateurs au classement des médecins de Midi-Pyrénées sur la base de leur montant de prescriptions pharmaceutiques. Cet article est le seul à ne pas contenir de données de durées.

Estimation de la fdr dans le cadre des sondages

Avec ma collègue Sandrine Casanova, enseignant-chercheur à UT1 Capitole, nous nous intéressons depuis plusieurs années à l'estimation de la fdr dans le cadre particulier des sondages lorsque la variable d'intérêt est censurée à droite. Dans [10], nous avons proposé un estimateur non paramétrique de la fdr en population finie avec une approche basée sur un modèle. Dans la prépublication [22], nous nous consacrons au même problème mais dans le cadre de l'estimation sur petits domaines.

2. Analyse de données censurées multivariées

Les données censurées multivariées, qui constituaient déjà le sujet de ma thèse de doctorat, se rencontrent très fréquemment, en particulier dans le domaine médical. On distingue le cas des événements répétés et celui des risques concurrents, où un seul événement, le premier, est observé.

Risques concurrents avec censure à droite

Avec Gérard Derzko, biostatisticien chez Sanofi-Synthélabo, nous avons montré dans [4] que dans le contexte des risques concurrents, les données observées pouvaient relever de deux mécanismes de création très différents, le mécanisme de sélection par minimum et le mécanisme de mélange censuré. Nous avons développé des algorithmes permettant d'estimer non paramétriquement les fonctions d'incidence cumulée d'intérêt dans ces deux cas distincts.

Événements répétés avec censure et événement terminal

Dans [5], toujours dans le cadre d'une collaboration avec Gérard Derzko, nous avons proposé un estimateur convergent non paramétrique de l'incidence cumulée d'événements répétés, rang par rang ou globale, en présence de censure à droite et de troncature des événements d'intérêt par un événement absorbant. L'estimation de la prévalence associée à un événement de rang donné s'en déduit simplement.

Suivi post-thérapeutique en oncologie

Un des domaines d'application des risques concurrents est l'oncologie, où le patient en rémission de son cancer est à risque de plusieurs types de rechutes : récidive du cancer sur le même site, récidive sur un autre site, métastases et décès.

Avec Thomas Filleron, méthodologiste-biostatisticien à l'Institut Claudius Regaud, centre anti-cancéreux situé à l'Oncopole de Toulouse, nous collaborons depuis plusieurs années sur divers sujets liés aux données censurées en lien avec l'oncologie. Avec notre étudiant de M2 puis de doctorat Serge Somda, qui a soutenu sa thèse en septembre 2015, nous nous sommes intéressés à la surveillance post-thérapeutique des patients en

oncologie, ce qui a donné lieu à plusieurs articles. Nous avons proposé dans [9] une méthode pour aider à la détermination de la durée optimale du suivi post-thérapeutique de patients en rémission de leur cancer en prenant en compte les risques concurrents associés aux différents types de rechute possibles. Une application à la détermination de la durée optimale du suivi dans le cas de tumeurs des cellules germinales, sans prendre en compte les risques concurrents, a fait l'objet de [13]. Dans [12], nous expliquons pourquoi notre approche de détermination de la durée de surveillance, basée sur les probabilités de rechute, doit être préférée à une approche basée sur les risques. Nous nous sommes ensuite penchés dans [11] sur le problème de la planification optimale des visites de contrôle lorsque la durée de surveillance et le nombre des visites sont fixés, en tenant toujours compte des risques concurrents et des facteurs pronostiques propres à chaque patient. Enfin, dans la prépublication [23], nous proposons un algorithme d'évaluation des différentes stratégies de surveillance.

3. Sélection de variables dans les modèles de survie en grande dimension

Comparaison de méthodes de sélection de modèles de survie

Avec Marie Walschaerts, alors étudiante en thèse sous la direction de Philippe Besse et Patrick Thonneau, nous nous sommes intéressées à la sélection de covariables dans des modèles où la variable réponse est censurée à droite. Nous avons recensé et comparé la stabilité de méthodes de sélection de variables basées sur le bootstrap pour deux méthodologies différentes communément utilisées dans les analyses de survie : le modèle des risques proportionnels de Cox et les arbres de survie. Cela a donné lieu à la prépublication [19].

Établissement d'une signature génomique en oncologie

A l'ère de la médecine personnalisée et avec les progrès du séquençage haut débit, une application récente en oncologie de la sélection de covariables est l'établissement de signatures génomiques, scores pronostiques basés sur l'expression des gènes et qui permettent de scinder les patients en deux groupes de bon et mauvais pronostic. Avec Marie Walschaerts et en collaboration avec des oncologues, nous avons établi dans [8] une signature génomique dans le cadre du cancer du poumon.

Sélection de modèles de survie en présence de risques concurrents

Enfin, un cas encore plus complexe est celui de la sélection de variables dans le cas des risques concurrents. Je me suis intéressée à ce problème dans le cadre d'un projet en collaboration avec divers centres anti-cancéreux et l'encadrement du stage de M2 de Soufiane Ajana. Nous avons comparé dans [14], sur des jeux de données réels et simulés, les performances de l'approche par *boosting* appliquée à un modèle à risques proportionnels et celle des forêts aléatoires de survie adaptées aux risques concurrents.

4. Autres travaux

Certains de mes travaux de recherche ne peuvent pas être classés dans les trois axes mentionnés ci-dessus. Je me contenterai de les décrire ici brièvement.

Écrit en collaboration avec des économistes, l'article [2] tente d'expliquer les variations des taux de chômage à un niveau sous-régional. Nous montrons sur des données

de la région Midi-Pyrénées que le meilleur modèle inclut une correction pour les erreurs spatialement auto-corrélées.

Dans [7] et [15], travaux menés en collaboration avec Jean-François Dupuy, alors enseignant-chercheur à l'Université Toulouse 3, nous avons considéré le cas encore jamais étudié d'un modèle de Cox stratifié où la variable de stratification est manquante pour certains sujets et avons étudié les propriétés asymptotiques de l'estimateur de calibration des données manquantes par régression dans ce contexte : bien que non convergent, il est asymptotiquement normal. Une étude de simulation a montré un biais modéré et une variance plus faible que celle de l'estimateur qui n'utilise que les cas complets.

Dans [17], qui fait suite aux travaux de thèse de Farah Salameh en Génie Electrique, nous comparons différents modèles de durées de vie d'isolants électriques, afin d'identifier les variables explicatives les plus influentes. En particulier, nous considérons une approche paramétrique avec des modèles de durées de vie accélérés que nous comparons à une approche non paramétrique basée sur des arbres et des forêts aléatoires de survie.

Le mémoire est organisé de la façon suivante. Après un bref rappel sur les méthodes de base de l'analyse des données censurées à droite dans le chapitre 1, les travaux des trois axes thématiques de recherche mentionnés ci-dessus seront décrits dans les trois chapitres suivants. Une conclusion générale sur mes perspectives de recherche clôt ce mémoire.

1 — Contexte et méthodes de base

Ce chapitre introduit le contexte des données censurées à droite et expose brièvement les techniques d'analyse statistique de base des données de survie, en particulier non paramétriques, auxquelles il sera fait référence dans la suite du mémoire. Pour un exposé plus complet sur l'analyse des données censurées, on peut par exemple consulter le livre de Klein et Moeschberger 1998 ou le livre de Com-Nougué et al. 1999 écrit en français par des enseignants-chercheurs de l'inserm.

1.1 Contexte

Dans le domaine de l'analyse des données censurées, appelées aussi données de survie, on s'intéresse au délai d'apparition d'un événement au cours du temps. Le premier événement qui a été enregistré est le décès par les démographes, ce qui explique que le terme de survie soit toujours employé de façon générale, bien que beaucoup d'autres événements que le décès soient maintenant objets d'étude. En effet, les données de survie ne sont pas l'apanage des biostatisticiens et sont aussi présentes dans des domaines comme la fiabilité, l'économie, l'assurance, la psychologie, la sociologie... En fiabilité industrielle, les événements d'intérêt sont par exemple les durées de vies des composants d'un système ; les économistes s'intéressent à des durées d'épisodes de chômage, les assureurs au temps d'arrivée d'un sinistre tandis que les psychologues mesurent le temps qu'il faut à un sujet pour accomplir une tâche donnée. En sociologie, le choix est vaste avec les successions des événements de vie : mariage, naissance du premier enfant, divorce...

Dans le domaine biomédical, les données de survie se rencontrent principalement en recherche clinique dans le cadre des essais thérapeutiques et dans les études de cohorte en épidémiologie. Dans le premier cas, on est souvent amené à comparer l'efficacité de deux traitements avec comme critère le délai jusqu'à un événement d'intérêt (rechute, décès...) ; dans le second cas, on suit au cours du temps une cohorte de sujets pour lesquels la maladie peut ou non apparaître pendant la période de suivi et l'on cherche à évaluer l'importance de facteurs de risque sur cette maladie.

La spécificité des durées de survie est de correspondre à des variables aléatoires positives et de comporter des observations incomplètes. En effet, dans la plupart des études prospectives, les individus sont suivis pendant une durée d'observation fixée à l'avance.

Pour les sujets pour lesquels l'événement d'intérêt a lieu pendant la période d'observation, on dispose du délai exact d'apparition de cet événement d'intérêt, mesuré depuis une date initiale qu'il faut spécifier sans ambiguïté (date de randomisation dans les essais cliniques, par exemple). Cependant, à la fin de la période d'observation, certains sujets n'auront pas eu l'événement. On n'aura alors pour ces sujets qu'une information incomplète, à savoir que le délai d'apparition de l'événement est plus grand que la durée d'observation. De telles données sont dites censurées à droite et la durée d'observation constitue le délai de censure. Comme les sujets rentrent dans l'étude à des dates différentes, chaque sujet a un délai d'apparition et un délai de censure qui lui sont propres. Le délai observé est donc le minimum du délai d'apparition de l'événement et du délai de censure. Ce mécanisme de censure constitue la censure aléatoire à droite, qui est le cas le plus courant dans les études prospectives et qui est le cas que nous considérons dans toute la suite. Les sujets qui n'ont pas eu l'événement d'intérêt pendant la période d'observation sont dits « exclus vivants » à la fin de l'étude [1]. Une autre cause de censure à droite, qu'on essaie de limiter au maximum dans les études, correspond aux sujets dits « perdus de vue », qui sont les individus qui ont quitté l'étude avant l'apparition de l'événement d'intérêt et dont on n'a plus de nouvelles à la date de fin d'étude.

D'autres types de censure peuvent être rencontrés : des données sont dites censurées à gauche si on ne connaît qu'une limite supérieure du délai d'apparition de l'événement ; elles sont censurées par intervalle si l'on sait seulement que l'événement a eu lieu entre deux dates connues (par exemple, passage du seuil critique d'un paramètre biologique entre deux visites du patient).

L'analyse des données censurées nécessite une méthodologie adaptée permettant de prendre en compte l'information contenue dans le délai de censure. Les procédures usuelles qui tiennent compte de la censure supposent presque toujours l'indépendance des variables durées de survie et délais de censure. La censure est de plus supposée la plupart du temps non informative, dans le sens où sa distribution ne dépend pas des paramètres qui interviennent dans la distribution de la variable durée de survie. Dans toute la suite, ces hypothèses sont conservées. Ces hypothèses sont raisonnables si les données censurées sont dues à des sujets qui n'ont pas expérimenté l'événement à la fin de l'étude (sujets « exclus vivants »). Elles sont moins évidemment vérifiées quand les données censurées correspondent à des sujets qui ont quitté l'étude avant la fin (« perdus de vue »).

1.2 Notations et fonctions d'intérêt

Nous choisissons d'utiliser les notations introduites par Cox 1972, plus intuitives et d'un formalisme plus léger que celles de l'approche par les processus ponctuels, que le lecteur intéressé pourra trouver dans Gill 1980 et Andersen, Borgan et al. 1993.

Nous notons T le délai jusqu'à l'événement d'intérêt. T est une variable aléatoire positive continue. Pour simplifier l'exposé de ces notions de base, on supposera que la variable T représente la durée de survie et donc que l'événement observé est le décès.

Les fonctions d'intérêt

Les fonctions les plus utilisées en analyse de la survie et qui caractérisent le mieux la distribution de T sont la fonction de survie S, la fonction de risque instantané h et la fonction de risque cumulé H.

1. Cette cause de censure à droite est parfois aussi appelée censure administrative.

La densité de probabilité de T est notée f, sa fonction de répartition est $F(t) = P(T \leq t)$ et sa fonction de survie est $S(t) = P(T > t) = 1 - F(t)$.

On définit le risque instantané de décès par

$$h(t) = \lim_{dt \to 0} \frac{P(t < T \leq t + dt \mid T > t)}{dt}.$$

$h(t)dt$ représente la probabilité de décéder entre t et $t + dt$ pour un sujet, conditionnellement au fait que ce sujet était encore vivant juste avant t.

La fonction h vérifie les relations suivantes :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \ln S(t).$$

La fonction de risque cumulé H est définie par

$$H(t) = \int_0^t h(s)ds$$

et elle est reliée à la fonction de survie par $H(t) = -\ln(S(t))$.

La censure à droite

La variable durée de survie T peut être censurée à droite par la variable Z, délai de censure, positive et continue. Il s'ensuit que le délai observé est le minimum de ces deux délais. Avec les notations d'Efron 1967, nous observons donc le couple (Y,Δ) où $Y = \min(T,Z)$ est le délai observé et $\Delta = \mathbb{1}(T \leq Z)$ est l'indicatrice d'événement. Le n-échantillon $(Y_i,\Delta_i)_{i=1,\ldots,n}$ constitue donc les données observées. Dans certains travaux, nous aurons aussi un p-vecteur de covariables que nous notons X_i.

Nous notons $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ les k temps de décès distincts non censurés observés ordonnés et $m_{(i)}$ le nombre de temps de décès égaux à $t_{(i)}$ ($m_{(i)} = 1$ s'il n'y a pas d'ex æquo). $\mathcal{R}(t_{(i)})$ désigne l'ensemble des sujets à risque de décès juste avant $t_{(i)}$, c'est-à-dire les sujets dont le délai observé Y_i est supérieur ou égal à $t_{(i)}$. $r_{(i)}$, cardinal de $\mathcal{R}(t_{(i)})$, est donc le nombre de sujets à risque de décès juste avant $t_{(i)}$.

1.3 Estimation non paramétrique des fonctions d'intérêt

L'estimateur le plus utilisé de la fonction de survie S, qui généralise l'estimateur empirique de la survie $1 - F_n$ au cas censuré, est l'estimateur proposé par Kaplan et Meier 1958 et défini par :

$$\widehat{S}_{\text{KM}}(t) = \begin{cases} 1 & \text{si } t < t_{(1)}, \\ \displaystyle\prod_{i:t_{(i)} \leq t} \left(1 - \frac{m_{(i)}}{r_{(i)}}\right) & \text{si } t \geq t_{(1)}. \end{cases} \tag{1.1}$$

Cet estimateur est une fonction en escalier décroissante avec des sauts en chaque valeur des temps de décès $t_{(i)}$. Il n'atteint 0 que si la plus grande observation $Y_{(n)}$ correspond à une observation non censurée.

L'estimateur de Kaplan-Meier converge presque sûrement et uniformément vers S (Földes et al. 1980). Sous certaines conditions de régularité, il converge en loi vers un processus gaussien (voir Breslow et Crowley 1974). Les propriétés mathématiques de

l'estimateur de Kaplan-Meier peuvent également être trouvées au chapitre 7 de Shorack et Wellner 1986.

De la relation $H(t) = -\ln(S(t))$, on peut dériver de l'estimateur de Kaplan-Meier un estimateur du risque cumulé H : cet estimateur $\hat{H}_{\mathrm{Br}}(t) = -\ln\left(\hat{S}_{\mathrm{KM}}(t)\right)$ est connu sous le nom d'estimateur de Breslow du risque cumulé.

Un estimateur plus connu du risque cumulé est celui de Nelson-Aalen (Nelson 1972 ; Aalen 1978), défini par :

$$\widehat{H}_{\mathrm{NA}}(t) = \begin{cases} 0 & \text{si } t < t_{(1)}, \\ \displaystyle\sum_{i:t_{(i)} \leq t} \frac{m_{(i)}}{r_{(i)}} & \text{si } t \geq t_{(1)}. \end{cases} \tag{1.2}$$

On peut à partir de $\widehat{H}_{\mathrm{NA}}$ obtenir un autre estimateur de la fonction de survie S d'après la relation $S(t) = \exp(-H(t))$, connu sous le nom d'estimateur de Harrington et Fleming.

1.4 Comparaison de la survie de deux ou plusieurs groupes

Les tests les plus utilisés pour comparer les fonctions de survie de deux ou plusieurs groupes sont les tests du logrank pondérés, qui sont des tests de rang.

L'hypothèse nulle H_0 correspond à l'égalité des fonctions de survie dans tous les groupes. Dans le cas de deux groupes A et B, les statistiques des tests du logrank pondérés s'écrivent

$$LR = \frac{\displaystyle\sum_{i=1}^{k} w_i \left(m_{B_{(i)}} - m_{(i)} \frac{r_{B_{(i)}}}{r_{(i)}} \right)}{\sqrt{\displaystyle\sum_{i=1}^{k} w_i^2 m_{(i)} \frac{r_{(i)} - m_{(i)}}{r_{(i)} - 1} \frac{r_{A_{(i)}} r_{B_{(i)}}}{r_{(i)}^2}}}, \tag{1.3}$$

où $m_{A_{(i)}}$ et $m_{B_{(i)}}$ sont les nombres de décès observés au temps $t_{(i)}$ dans chacun des groupes A et B et $r_{A_{(i)}}$ et $r_{B_{(i)}}$ les nombres de sujets exposés au risque de décès juste avant $t_{(i)}$. Les statistiques des tests du logrank pondérés suivent asymptotiquement sous H_0 une loi normale centrée réduite.

Les poids w_i permettent de donner plus d'importance à certains décès. Si tous les poids sont égaux à 1, on a la pondération de Mantel-Haenszel correspondant au test appelé plus simplement test du logrank. Des poids décroissants avec $t_{(i)}$ donnent plus de poids aux décès précoces. De façon générale, il a été montré (voir Gill 1980) que les poids asymptotiquement optimaux pour une alternative donnée sont proportionnels à $\ln\left(\frac{h_B(t)}{h_A(t)}\right)$, où h_A et h_B sont les risques instantanés de décès dans les groupes A et B respectivement. Ainsi le test du logrank (avec des poids constants égaux à 1) est optimal pour les alternatives à risques proportionnels, qui correspondent à l'hypothèse fondamentale du modèle de Cox.

1.5 Le modèle à risques proportionnels de Cox

Cox 1972 a proposé de modéliser le risque instantané de décès sachant les variables explicatives par un modèle semi-paramétrique qui présente l'avantage de ne faire aucune

hypothèse sur la distribution de la variable durée de survie T. Ce modèle est vite devenu le modèle de régression pour données de survie le plus utilisé, en particulier dans le domaine médical. Dans le modèle de régression de Cox sous sa forme la plus simple, la fonction de risque instantané de décès sachant les covariables est modélisée par le produit suivant :

$$h(t|X) = h_0(t)\exp(\beta' X), \tag{1.4}$$

où X est un p-vecteur de covariables, β est le p-vecteur des paramètres de régression et h_0 est une fonction de risque de base inconnue.

Le modèle de Cox est un modèle log-linéaire, basé sur l'hypothèse des risques proportionnels, qui stipule que le rapport des risques instantanés de deux individus est constant au cours du temps et ne dépend donc que de leurs covariables. En effet, d'après le modèle 1.4, le rapport des risques instantanés pour deux sujets de covariables X_1 et X_2 ne dépend que des covariables et non du temps. On a :

$$\frac{h(t,X_1)}{h(t,X_2)} = \frac{\exp(\beta' X_1)}{\exp(\beta' X_2)} = \exp\left(\beta'(X_1 - X_2)\right).$$

Cox 1972 a proposé de façon heuristique d'estimer le vecteur des paramètres β et d'effectuer des tests de significativité en maximisant ce qu'il a appelé la vraisemblance partielle, en considérant la fonction de risque de base h_0 comme un paramètre de nuisance. Ultérieurement, Gill 1980 en a démontré rigoureusement les propriétés asymptotiques dans le cadre de la théorie des processus ponctuels, ce qui permet d'utiliser la vraisemblance partielle de Cox comme une fonction de vraisemblance usuelle (tests de Wald, du score ou du rapport de vraisemblance sur les paramètres du modèle).

La vraisemblance partielle de Cox s'écrit comme un produit de vraisemblances conditionnelles calculées en chaque temps $t_{(i)}$, les $t_{(i)}$ étant supposés fixés. La contribution V_i à la vraisemblance du sujet i (de vecteur de covariable X_i) dont le décès a été observé en $t_{(i)}$ est égale à la probabilité, conditionnelle à $\mathcal{R}(t_{(i)})$, que ce soit précisément ce sujet qui décède en $t_{(i)}$ parmi l'ensemble $\mathcal{R}(t_{(i)})$ des sujets à risque en $t_{(i)}$. On a donc

$$V_i = \frac{h(t_{(i)},X_i)}{\sum_{j\in\mathcal{R}(t_{(i)})} h(t_{(i)},X_j)}.$$

Sous le modèle de Cox, on vérifie que le terme de nuisance $h_0(t_{(i)})$ s'élimine dans cette expression et il reste

$$V_i = \frac{\exp(b' X_i)}{\sum_{j\in\mathcal{R}(t_{(i)})} \exp(b' X_j)}.$$

La vraisemblance partielle de Cox se calcule alors comme le produit sur i de toutes les contributions des sujets décédés et en l'absence de durées de vie ex æquo, la fonction de log-vraisemblance partielle vaut alors

$$L(\beta) = \sum_{i=1}^{k} \left(\beta' X_i - \ln\left(\sum_{j\in\mathcal{R}(t_{(i)})} e^{\beta' X_j} \right) \right). \tag{1.5}$$

Différents moyens pour prendre en compte les durées de vie ex æquo dans la log-vraisemblance partielle ont été proposées dans la littérature (voir en particulier Breslow 1974 et Efron 1977).

Le modèle de Cox a donné lieu à de nombreuses extensions, dont on pourra trouver un bon aperçu dans l'ouvrage de Therneau et Grambsch 2000.

1.6 Références

Aalen, O. (1978). « Nonparametric Inference for a Family of Counting Processes ». In : *The Annals of Statistics* 6.4, p. 701–726. URL : http://www.jstor.org/stable/2958850.

Andersen, P. K., O. Borgan, R. D. Gill et N. Keiding (1993). *Statistical Models Based on Counting Processes*. New-York : Springer-Verlag.

Breslow, N. (1974). « Covariance Analysis of Censored Survival Data ». In : *Biometrics* 30.1, p. 89–99. URL : http://www.jstor.org/stable/2529620.

Breslow, N. et J. Crowley (1974). « A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship ». In : *Annals of Statistics* 2, p. 437–453. DOI : 10.1214/aos/1176342705.

Com-Nougué, C., C. Hill, C. Kramar et T. Moreau (1999). *Analyse statistique des données de survie*. Statistique en Biologie et en Medecine. Médecine Sciences Publications. ISBN : 2257123107.

Cox, D. R. (1972). « Regression Models and Life-Tables ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2, p. 187–220. URL : http://www.jstor.org/stable/2985181.

Efron, B. (1967). « The Two Sample Problem with Censored Data ». In : *Proc. 5th Berkeley Symp.* 4, p. 831–853.

— (1977). « The Efficiency of Cox's Likelihood Function for Censored Data ». In : *Journal of the American Statistical Association* 72.359, p. 557–565. URL : http://www.jstor.org/stable/2286217.

Földes, A., L. Rejto et B. B. Winter (1980). « Strong consistency properties of nonparametric estimators for randomly censored: I: the product-limit estimator ». In : *Periodica Mathematica Hungarica* 11.3, p. 233–250.

Gill, R. D. (1980). « Censoring and Stochastic Integrals ». In : *Statistica Neerlandica* 34.2, p. 124–124. URL : http://dx.doi.org/10.1111/j.1467-9574.1980.tb00692.x.

Kaplan, E. L. et P. Meier (1958). « Nonparametric Estimation from Incomplete Observations ». In : *Journal of the American Statistical Association* 53.282, p. 457–481. URL : http://www.jstor.org/stable/2281868.

Klein, J. P. et M. L. Moeschberger (1998). *Survival Analysis - Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer. ISBN : 978-0-387-21645-4. URL : http://www.springer.com/gp/book/9780387953991.

Nelson, W. (1972). « Theory and Applications of Hazard Plotting for Censored Failure Data ». In : *Technometrics* 14.4, p. 945–966. URL : http://www.jstor.org/stable/1267144.

Shorack, G. R. et J. A. Wellner (1986). *Empirical Processes with Applications to Statistics*. John Wiley & Sons, New-York.

Therneau, T. M. et P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer. ISBN : 978-1-4757-3294-8.

2 — Estimation de la fonction de répartition

Il est fréquent d'avoir à estimer la fonction de répartition (fdr) d'une variable aléatoire T, éventuellement censurée à droite, afin de pouvoir en déduire des fonctionnelles d'intérêt comme les quantiles. Des fonctionnelles d'intérêt moins connues sont les expectiles, qui permettent de tenir compte des valeurs observées et pas seulement de leur rang. Dans certains cas, on cherche à avoir des estimateurs de la fdr conditionnellement à la valeur prise par une covariable X, afin d'en déduire par exemple les quantiles conditionnels de T sachant X. L'utilisation de tels modèles permet de construire des courbes de référence ou de comparer des groupes. Ces problèmes sont fréquemment rencontrés en médecine (G. Li et al. 1996), en psychologie, en économie et en fiabilité, où l'on peut par exemple vouloir estimer la médiane de vie d'un appareil soumis à un stress fixé.

Comme il est souvent difficile d'avoir une idée de la loi de T, nous avons privilégié l'approche non paramétrique pour tous les estimateurs proposés. L'estimateur non paramétrique le plus utilisé de la fdr d'une variable T censurée à droite est obtenu à partir de l'estimateur de Kaplan-Meier de la fonction de survie (Kaplan et Meier 1958, voir formule 1.1). Dans le cas où l'on s'intéresse à l'estimation de la fdr conditionnellement à une covariable X, plusieurs généralisations de l'estimateur de Kaplan-Meier ont été proposées (voir Beran 1981 ; Gonzalez-Manteiga et Cadarso-Suarez 1994) mais, à l'instar de l'estimateur de Kaplan-Meier, elles ne sont lisses que par rapport à la covariable X.

Dans la section 2.1, nous présentons l'article [3] et montrons comment le fait de lisser par rapport au temps des estimateurs de la fdr conditionnelle permet d'améliorer les performances des estimateurs, à la fois pour la fdr mais aussi pour les estimateurs des fonctions quantiles qui en découlent. Dans la section 2.2, nous nous intéressons à l'estimation d'ordres-expectiles proposée dans l'article [6]. A noter que cette section est la seule à ne pas comporter de données de survie. Enfin, dans la section 2.3, nous considérons le cas de l'estimation de la fdr dans le cadre particulier des sondages, d'abord pour une population finie globale (article [10]) puis dans le cadre des petits domaines (article [22]).

2.1 Estimation de la fdr conditionnelle et de ses quantiles

2.1.1 Motivation

Nous considérons un modèle de régression non paramétrique dans lequel le couple aléatoire (T,X) est tel que la variable T correspond à une durée de survie qui peut être censurée à droite par une variable aléatoire Z indépendante de T conditionnellement à la covariable X supposée continue. On observe donc ici $Y = \min(T,Z)$ et l'indicatrice d'événement $\Delta = \mathbb{1}(T \leq Z)$. Notre but est d'estimer les quantiles de la fdr conditionnelle de T sachant X dans une optique non paramétrique.

Dans le cas non censuré, une grande variété d'estimateurs non paramétriques des fonctions quantiles ont été proposés dans la littérature (voir Poiraud-Casanova et Thomas-Agnan 1998). Dans le cas où la variable réponse est censurée, une famille d'estimateurs qui généralisent l'estimateur de Kaplan-Meier de la fonction de survie (Kaplan et Meier 1958) a été proposée pour estimer la fdr conditionnelle dans les modèles avec covariable (voir Beran 1981 ; Gonzalez-Manteiga et Cadarso-Suarez 1994). Ces estimateurs font intervenir des poids : Dabrowska 1992 utilise des poids de type Nadaraya-Watson alors que Van Keilegom et Veraverbeke 1996 utilisent des poids de type Gasser-Müller. Ces derniers auteurs en déduisent un estimateur de la fonction quantile par inversion de l'estimateur de la fdr conditionnelle. Ces estimateurs sont lisses par rapport à la covariable mais ne le sont pas par rapport à l'ordre α du quantile alors que la fonction quantile théorique est une fonction continue de α. D'autre part, dans les modèles sans covariable, des estimateurs de la fonction quantile lisses par rapport à α ont été proposés (Padgett 1986 ; Padget et Thombs 1988).

En collaboration avec Sandrine Casanova et Christine Thomas-Agnan, j'ai donc proposé dans [3] des estimateurs de la fdr conditionnelle doublement lisses, c'est-à-dire lisses par rapport au temps et par rapport à la covariable X. Les estimateurs de la fonction quantile conditionnelle qui en découlent sont donc doublement lisses par rapport à α et la covariable. Ils seront présentés dans la section 2.1.2. Le bénéfice attendu de ce double lissage est une diminution de l'erreur quadratique moyenne (Falk 1983 ; Falk 1984), que nous vérifions par simulation dans la section 2.1.3. Une application sur des données médicales est présentée à la section 2.1.4 puis quelques extensions sont envisagées en section 2.1.5.

2.1.2 Les estimateurs doublement lisses

Nous avons proposé deux classes d'estimateurs, que nous allons détailler après avoir défini les notations.

Notations

L'échantillon observé est $(Y_i,\Delta_i,X_i)_{i=1}^n$ où $Y_i = \min(T_i,Z_i)$ et $\Delta_i = \mathbb{1}(T_i \leq Z_i)$. I désigne l'ensemble des indices des observations non censurées : $I = \{i : \Delta_i = 1\}$ et pour $i \in I$, nous notons $Y_i^\dagger = Y_i$. Nous utilisons de plus les conventions suivantes : $Y_{(0)}^\dagger = 0$ et $Y_{(\#I+1)}^\dagger = Y_{(n)}$ où les parenthèses indiquent les statistiques d'ordre. Pour un noyau de densité K, nous notons H le noyau intégré associé défini par $H(t) = \int_{-\infty}^t K(u)\ du$.

La première classe d'estimateurs

La première classe d'estimateurs est obtenue en lissant par rapport au temps les estimateurs de Kaplan-Meier généralisés de la fdr conditionnelle proposés par Beran 1981 et Gonzalez-Manteiga et Cadarso-Suarez 1994.

Les estimateurs de Kaplan-Meier généralisés de la fdr conditionnelle s'écrivent : $\widehat{F}_{\text{GKM}}(t \mid x) = 1 - \widehat{S}_{\text{GKM}}(t \mid x)$, où $\widehat{S}_{\text{GKM}}(t \mid x)$ est un estimateur de Kaplan-Meier géné-

ralisé de la fonction de survie, défini par

$$\widehat{S}_{\text{GKM}}(t \mid x) = \begin{cases} \displaystyle\prod_{i=1}^{n} \left\{ 1 - \frac{B_i(x)}{\displaystyle\sum_{r=1}^{n} \mathbb{1}(Y_r \geq Y_i) B_r(x)} \right\}^{\mathbb{1}(Y_i \leq t, \Delta_i = 1)} & \text{si } \; t < Y_{(n)} \\ 0 & \text{sinon,} \end{cases} \tag{2.1}$$

où les $B_i(x)$ $(i = 1, \ldots, n)$ sont des poids positifs et de somme 1. Nous avons utilisé les poids de type Nadaraya-Watson proposés par Dabrowska 1992 et définis par

$$B_i(x) = \frac{K\left(\dfrac{x - X_i}{h_X}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_X}\right)}, \tag{2.2}$$

où h_X est une fenêtre adaptée.

Il est facile de vérifier que $\widehat{F}_{\text{GKM}}$ est une fdr. Sa convergence uniforme presque sûre a été montrée par Dabrowska 1989. Par la suite, Gonzalez-Manteiga et Cadarso-Suarez 1994 ont établi pour cet estimateur un résultat de normalité asymptotique avec un facteur de normalisation égal à $\sqrt{nh_X}$.

Nous proposons d'obtenir une version lisse par rapport à t de ces estimateurs à l'aide d'un noyau intégré H :

$$\widehat{F}_{\text{SGKM}}(t \mid x) = \sum_{i=1}^{\#I+1} \left(\widehat{F}_{\text{GKM}}\left(Y^{\dagger}_{(i)} \mid x\right) - \widehat{F}_{\text{GKM}}\left(Y^{\dagger}_{(i-1)} \mid x\right) \right) H\left(\frac{t - Y^{\dagger}_{(i)}}{h_T}\right) \tag{2.3}$$

où h_T désigne une fenêtre appropriée. On peut noter que ce lissage est similaire au lissage classique par noyau de la fdr empirique (voir Prakasa Rao 1983) où les sauts en $\frac{1}{n}$ de la fdr empirique sont remplacés par les sauts de l'estimateur de Kaplan-Meier généralisé. On peut facilement vérifier que $\widehat{F}_{\text{SGKM}}$ est une fdr.

L'estimation de la fonction quantile conditionnelle $\widehat{q}_{\text{SGKM}}(\alpha,x)$, où α $(0 \leq \alpha \leq 1)$ désigne l'ordre du quantile, s'obtient ensuite simplement par inversion numérique de $\widehat{F}_{\text{SGKM}}(\cdot \mid x)$. Comme attendu, cette fonction est lisse par rapport à α et x.

La deuxième classe d'estimateurs

Une seconde approche consiste à lisser conjointement par rapport à t et à la covariable x. Nous estimons d'abord la fdr jointe du couple (T,X). Comme seule T est censurée, nous pouvons tirer partie de cette asymétrie en écrivant :

$$F_{T,X}(t,x) = \mathrm{P}(T \leq t, X \leq x) = \mathrm{P}(T \leq t \mid X \leq x)\, \mathrm{P}(X \leq x). \tag{2.4}$$

Le premier facteur du produit est estimé par l'estimateur de Kaplan-Meier appliqué au sous-échantillon de données pour lesquelles les valeurs de la covariable sont inférieures ou égales à x. Le second facteur est estimé par la fdr empirique de X, notée $\widehat{F}_X$. Par construction, l'estimateur obtenu, noté $\widehat{F}(t,x)$, est une fonction en escalier par rapport à t et à x que nous pouvons écrire ainsi :

$$\widehat{F}(t,x) = \sum_{i=1}^{\#I+1} \sum_{j=1}^{n} \omega_{ij} \mathbb{1}\left(Y^{\dagger}_{(i)} \leq t, X_{(j)} \leq x\right), \tag{2.5}$$

avec les poids ω_{ij} qui correspondent à des sauts bivariés (cf. [3] pour le calcul de ces poids). Ces sauts peuvent être négatifs, si bien que $\widehat{F}$ n'est pas une fonction de répartition.

La seconde étape est celle du lissage conjoint. Nous nous inspirons de Samanta 1989 et de Berlinet et al. 1998 qui, dans le cas non censuré, ont défini un estimateur de la fdr conditionnelle de T sachant $X = x$ par

$$\widehat{F}(t \mid x) = \int_{-\infty}^{t} \frac{\widehat{f}_{T,X}(u,x)}{\widehat{f}_X(x)}\, du = \frac{1}{\widehat{f}_X(x)} \int_{-\infty}^{t} \widehat{f}_{T,X}(u,x)\, du, \tag{2.6}$$

où $\widehat{f}_X$ et $\widehat{f}_{T,X}$ sont respectivement les estimateurs à noyau de Parzen-Rosenblatt de la fonction de densité marginale de X et de la fonction de densité jointe de T et de X.

Dans notre cas où T est censuré, nous ne pouvons pas utiliser l'estimateur usuel bivarié de Parzen-Rosenblatt et nous utilisons l'estimateur $\widehat{F}$ obtenu à la première étape pour proposer l'estimateur de la densité jointe suivant :

$$\widehat{f}_{T,X}(u,x) = \sum_{i=1}^{\#I+1} \sum_{j=1}^{n} \omega_{ij} (h_T h_X)^{-1} K\left(\frac{u - Y_{(i)}^{\dagger}}{h_T}\right) K\left(\frac{x - X_{(j)}}{h_X}\right),$$

où h_T et h_X sont deux fenêtres appropriées.

En adaptant (2.6), nous obtenons finalement l'expression ci-dessous pour l'estimateur de la fdr conditionnelle avec lissage conjoint :

$$\widehat{F}_{\text{JS}}(t \mid x) = \frac{n \displaystyle\sum_{i=1}^{\#I+1} \sum_{j=1}^{n} \omega_{ij} K\left(\frac{x - X_{(j)}}{h_X}\right) H\left(\frac{t - Y_{(i)}^{\dagger}}{h_T}\right)}{\displaystyle\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_X}\right)}. \tag{2.7}$$

Une version monotone croissante (notée $\widehat{F}_{\text{IJS}}$) peut être obtenue et il peut être montré très simplement que nous obtenons alors une fdr. Les fonctions quantiles s'obtiennent par inversion numérique comme pour la première classe d'estimateurs.

Lien entre les deux classes d'estimateurs

Les deux classes d'estimateurs sont construites de façon très différente et dans le cas général il est difficile de les comparer sur un plan théorique. Cependant, dans le cas où T n'est pas censurée, nous avons la propriété suivante, dont la démonstration est en annexe de l'article [3].

Proposition 2.1.1 Si T n'est pas censuré et si les poids de $\widehat{F}_{\text{SGKM}}(\cdot \mid x)$ sont les poids de type Nadaraya-Watson $B_i(x) = \dfrac{K\left(\dfrac{x - X_i}{h_X}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\dfrac{x - X_i}{h_X}\right)}$, alors $\widehat{F}_{\text{SGKM}}(\cdot \mid x) = \widehat{F}_{\text{JS}}(\cdot \mid x)$ et les deux estimateurs sont égaux à l'estimateur de la fdr conditionnelle défini par Samanta 1989 et Berlinet et al. 1998.

Un corollaire immédiat de cette proposition est que si T n'est pas censurée, alors $\widehat{F}_{\text{JS}}$ est monotone et on a alors $\widehat{F}_{\text{IJS}} = \widehat{F}_{\text{JS}}$.

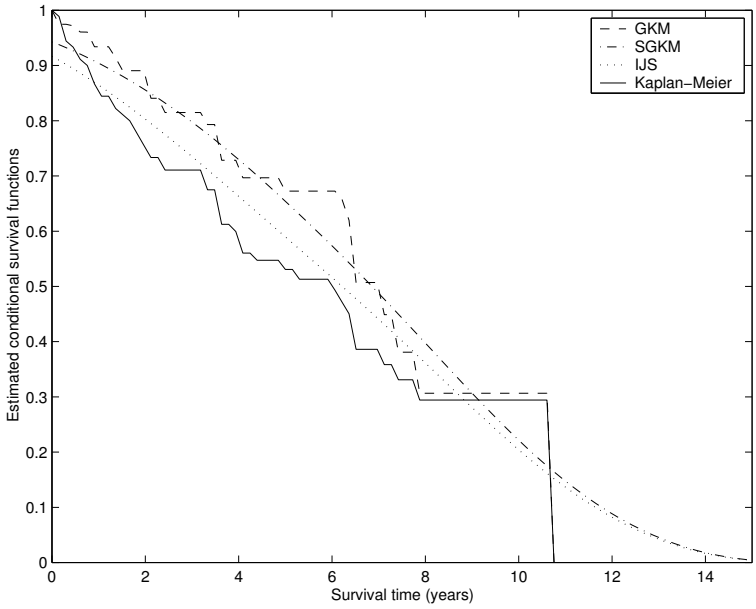


FIGURE 2.1 – Estimateurs de la fonction de survie conditionnelle pour $x = 65$ ans.

2.1.3 Simulations

Des simulations ont été effectuées pour comparer les performances des nouveaux estimateurs de la fdr conditionnelle à celles des estimateurs de Kaplan-Meier généralisés, pour différentes tailles d'échantillons et différents taux de censure. Sur la base du critère MASE (Mean Averaged Squared Error), comme attendu, l'estimation est nettement améliorée lorsqu'on lisse par rapport aux deux variables. Les deux nouvelles classes d'estimateurs montrent des performances quasi similaires, avec un léger avantage pour la deuxième classe. Les mêmes résultats sont obtenus en ce qui concerne les estimateurs correspondants des quantiles conditionnels.

2.1.4 Application

A titre d'illustration, nous avons appliqué nos méthodes aux données de KARDAUN 1983, qui concernent 90 patients pour lesquels un cancer du larynx a été diagnostiqué entre 1970 et 1978. L'événement étudié est le décès du patient. Le taux de censure est de 45 %. Nous nous intéressons au délai de survie en fonction de l'âge au diagnostic.

La figure 2.1 montre les courbes des trois estimateurs $\widehat{S}_{\text{GKM}}$, $\widehat{S}_{\text{SGKM}}$ et $\widehat{S}_{\text{IJS}}$ de la fonction de survie conditionnelle évaluée à l'âge médian (65 ans). L'estimateur de Kaplan-Meier de la fonction de survie globale a été tracé pour référence. On peut voir sur la figure 2.2 les trois estimateurs des quantiles conditionnels $\widehat{q}_{\text{GKM}}$, $\widehat{q}_{\text{SGKM}}$ et $\widehat{q}_{\text{IJS}}$ en fonction de α. On peut noter la proximité des estimateurs $\widehat{S}_{\text{SGKM}}$ and $\widehat{S}_{\text{IJS}}$ ainsi que celle des estimateurs correspondants des quantiles $\widehat{q}_{\text{SGKM}}$ et $\widehat{q}_{\text{IJS}}$.

La figure 2.3 présente l'estimateur $\widehat{S}_{\text{IJS}}$ calculé pour les trois quartiles de l'âge, ce qui permet aux médecins d'étudier le rôle de l'âge au diagnostic sur la survie du patient. Enfin, la figure 2.4 montre les courbes de l'estimateur $\widehat{q}_{\text{IJS}}$ du quantile conditionnel en fonction de l'âge pour différentes valeurs de α.

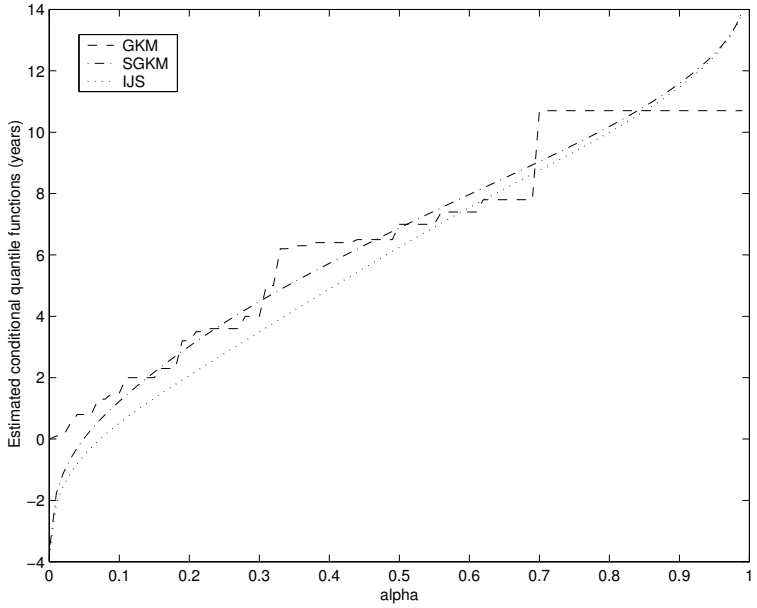


FIGURE 2.2 – Estimateurs de la fonction quantile conditionnelle pour $x = 65$ ans.

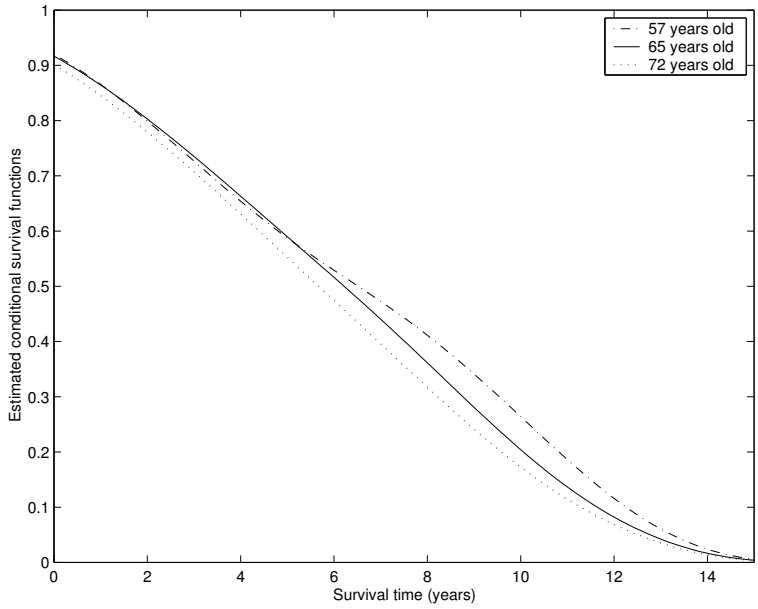


FIGURE 2.3 – Estimateurs $\widehat{S}_{\text{IJS}}$ de la fonction de survie conditionnelle pour les quartiles de l'âge au diagnostic.

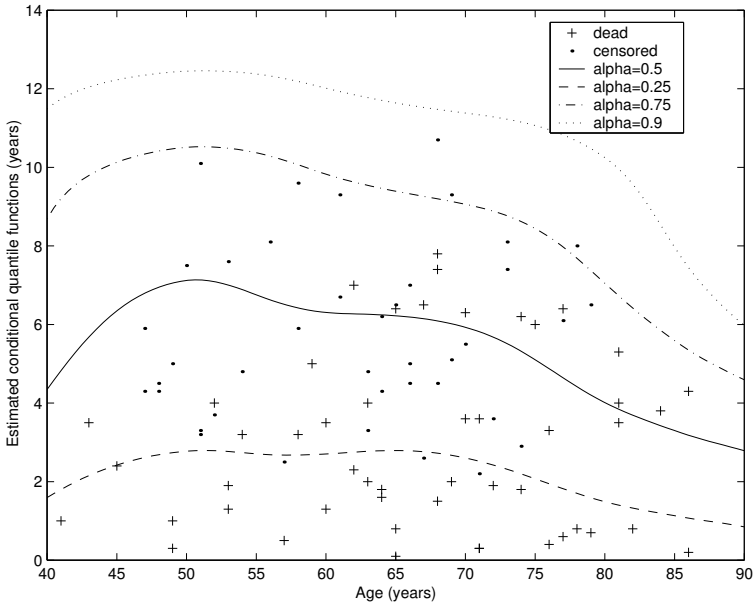


FIGURE 2.4 – Estimateurs $\widehat{q}_{\text{IJS}}$ de la fonction quantile conditionnelle en fonction de l'âge au diagnostic pour différentes valeurs de α.

2.1.5 Extensions et perspectives

Cas d'une covariable censurée

Nos méthodes peuvent facilement s'étendre au cas d'une covariable censurée. En effet, pour la deuxième classe d'estimateurs, l'estimateur $\widehat{F}$ de la fdr bivariée peut se généraliser au cas où X est censurée à droite en utilisant les résultats de CAMPBELL et FÖLDES 1982. Le terme $P(X \leq x)$ peut être estimé par l'estimateur de Kaplan-Meier de la fonction de survie marginale. De plus, si nous notons C le délai de censure aléatoire associé à la covariable X, V le minimum de X et de C et si nous supposons que (T,X) et (Z,C) sont mutuellement indépendants, alors le terme $\mathrm{P}(T \leq t \mid X \leq x)$ est égal à $\Big[\mathrm{P}(T \leq t) - \mathrm{P}(X > x)\left\{1 - \mathrm{P}(T > t \mid V > x)\right\}\Big] / \mathrm{P}(X \leq x)$.

Ainsi, $\mathrm{P}(T \leq t)$ peut être estimé par l'estimateur de Kaplan-Meier de la fonction de survie marginale et le terme $\mathrm{P}(T > t \mid V > x)$ peut être estimé par l'estimateur de Kaplan-Meier calculé sur le sous-échantillon des données pour lesquelles la valeur observée de V est plus grande que x.

Perspectives

D'autres types de lissages pourraient être envisagés. Le lissage de la première étape pourrait être effectué en utilisant des polynômes locaux ou des splines de lissage. Quand à la deuxième étape, qui consiste à lisser une fdr discrète par rapport à la variable réponse, cela pourrait également être réalisé à l'aide de splines de lissage. Pour ce qui est de la seconde classe d'estimateurs, on peut penser à obtenir un estimateur de la fdr bivariée à partir d'un estimateur de la fonction de risque cumulé bivariée (voir CAMPBELL 1981). Cela devrait permettre un meilleur contrôle des effets de bord. Une autre piste d'amélioration consisterait à étudier des techniques de choix de fenêtres basées sur les données adaptées à nos estimateurs. Enfin, on pourrait penser à obtenir un estimateur lisse de la fonction quantile par une méthode directe, c'est-à-dire sans inverser un estimateur de la fdr conditionnelle. Dans le cas non censuré, des méthodes utilisant des

splines de régression, des splines de lissage ou des polynômes locaux se sont révélées efficaces (voir par exemple Poiraud-Casanova et Thomas-Agnan 1998). Il faudrait étudier la possibilité de les adapter au cas censuré.

2.2 Estimation d'ordres-expectiles

2.2.1 Motivation

Si le plus souvent les méthodes statistiques visent à étudier des comportements moyens d'individus, dans certains cas, il est plus pertinent de s'intéresser à des comportements extrêmes, caractérisés par les valeurs extrêmes d'une certaine mesure de performance. On peut penser par exemple à l'efficacité d'agences bancaires ou aux performances commerciales d'un réseau de revendeurs. Il est naturel de tenter d'identifier ces individus extrêmes à partir d'un classement de ces individus.

Considérons le cas où le résultat d'un individu est mesuré par une variable Y et où les variables explicatives correspondantes sont mesurées par un vecteur de covariables X. La recherche de comportements extrêmes se décompose habituellement en deux étapes : d'abord la recherche d'un comportement moyen par une technique de régression, suivie d'une mesure de l'écart (ou performance) de chaque observation à ce comportement moyen, basée par exemple sur l'écart-type de l'erreur d'ajustement. Cette méthode peut se révéler assez lourde. Elle suppose de plus assez souvent une hypothèse de normalité des erreurs qui peut ne pas être réaliste. Un autre problème majeur en construisant un tel classement réside dans l'hétéroscédasticité des erreurs de la régression. En particulier, il se peut que les valeurs observées pour les individus qui s'éloignent de la moyenne reflètent simplement l'hétéroscédasticité induite par les variables explicatives et non le comportement intrinsèque de ces individus. Cette hétéroscédasticité est le plus souvent prise en compte en mettant en œuvre une régression pondérée. Cependant, une telle approche nécessite, en plus des hypothèses paramétriques sur la fonction de régression, de faire des spécifications pour modéliser l'hétéroscédasticité. Enfin, les erreurs sont souvent supposées distribuées de façon symétrique, ce qui peut ne pas être le cas.

Des approches non paramétriques ont été proposées pour ajuster des modèles hétéroscédastiques (voir par exemple Welsh 1996) mais elles peuvent s'avérer complexes. Une autre solution est de traiter le problème par une approche directe qui modélise les quantiles conditionnels de la réponse sachant les covariables. Rappelons que les quantiles font partie de la classe plus générale des fonctionnelles de position des distributions que Breckling et Chambers 1988 ont dénommées les M-quantiles. Une classe moins connue que les quantiles est la classe des expectiles, qui généralisent l'espérance de la même façon que les quantiles généralisent la médiane (Newey et Powell 1987).

Notre motivation pour ce problème est issue de données de l'Union Régionale des Caisses d'Assurance Maladie (urcam) relatives aux médecins de la région Midi-Pyrénées en 1999. L'urcam souhaitait ordonner ces médecins sur la base de leur comportement de prescription de médicaments en tenant compte des caractéristiques de leur pratique médicale ainsi que d'autres variables pertinentes. Ce classement peut permettre d'identifier des médecins ayant un comportement de prescription particulier mais aussi de repérer s'il y a des groupements des ces rangs extrêmes associés à une sous-région donnée, ce qui pourrait indiquer des inégalités locales en termes de dépenses de santé.

Si on suppose qu'un médecin ayant un montant de prescription moyen sur une période donnée qui dépasse un certain seuil, noté y_0, génère une « perte » pour la Sécurité Sociale, alors la perte moyenne par médecin vaut :

$$E((Y - y_0)\mathbb{1}(Y > y_0)) \tag{2.8}$$

tandis que la probabilité qu'un médecin dépasse le seuil critique vaut

$$E(\mathbb{1}(Y > y_0)). \tag{2.9}$$

D'un point de vue économique, la Sécurité Sociale est plus intéressée par (2.8) que par (2.9). De plus, il est clair que le montant de prescription d'un médecin dépend de ses caractéristiques personnelles et de son environnement, si bien que le seuil y_0 va aussi en dépendre. Ce seuil y_0 n'est pas connu mais il possible d'utiliser l'argument ci-dessus pour classer les praticiens. Considérons un médecin avec un montant de prescription $Y = y_i$ et un vecteur de covariables $\mathrm{X} = \mathrm{x}_i$. La perte additionnelle attendue par la Sécurité Sociale causée par une augmentation de la prescription moyenne de ce médecin vaut

$$E\left((Y - y_i)\mathbb{1}(Y > y_i) \mid \mathrm{X} = \mathrm{x}_i\right). \tag{2.10}$$

Pour obtenir un coefficient normalisé, il suffit de le diviser par l'écart absolu moyen : $E\left((|Y - y_i| \mid \mathrm{X} = \mathrm{x}_i\right)$. On obtient alors :

$$\frac{E(|Y - y_i|\mathbb{1}(Y > y_i) \mid \mathrm{X} = \mathrm{x}_i)}{E(|Y - y_i| \mid \mathrm{X} = \mathrm{x}_i)}. \tag{2.11}$$

Plus ce rapport est grand, moins le médecin a un comportement de prescription risqué pour la Sécurité Sociale (voir Newey et Powell 1987). En effet, étant donné ses caractéristiques, on attend de lui un montant de prescription additionnel plus élevé. Nous allons donc, pour pouvoir associer un classement élevé avec un risque élevé, considérer le rapport complémentaire suivant

$$q_i = \frac{E(|Y - y_i|\mathbb{1}(Y \leq y_i) \mid \mathrm{X} = \mathrm{x}_i)}{E(|Y - y_i| \mid \mathrm{X} = \mathrm{x}_i)} \tag{2.12}$$

que nous appelons « ordre-expectile » du comportement de prescription du médecin [1].

Cette définition de l'ordre-expectile fait pendant à celle de l'ordre-quantile, défini par

$$\frac{E(\mathbb{1}(Y \leq y_i) \mid \mathrm{X} = \mathrm{x}_i)}{E(1 \mid \mathrm{X} = \mathrm{x}_i)} \tag{2.13}$$

qui correspond à la probabilité qu'un médecin de caractéristiques x_i ait un montant de prescription inférieur ou égal à y_i. L'utilisation d'un classement basé sur les ordres-expectiles se justifie donc ici parce que le montant des dépassements importe plus que leur rang pour le budget de la Sécurité Sociale.

La méthode de classement que nous avons proposée dans [6] est donc basée sur la régression expectile, dont nous rappelons le principe dans la section 2.2.2.

Nous proposons d'estimer les expectiles conditionnels par les méthodes de la constante locale et des polynômes locaux de degré 1 pour une grille de valeurs d'ordres et nous en déduisons l'ordre-expectile par interpolation linéaire. L'estimation par polynômes locaux ne donnant pas nécessairement un estimateur monotone par rapport à q, nous avons envisagé différentes techniques de monotonisation (Mukarjee et Stern 1994). Alternativement, nous avons proposé une technique d'estimation directe de l'ordre-expectile en utilisant des estimateurs à noyau de type Nadaraya-Watson calibrés (Hall et al. 1999). Ces méthodes seront détaillées dans la section 2.2.3. Les propriétés à distance finie de nos méthodes ont été comparées par simulation (voir section 2.2.4) puis appliquées sur les données de prescription des médecins de Midi-Pyrénées, présentées à la section 2.2.5. La section 2.2.6 envisagera quelques perspectives.

1. Pour ne pas perdre le lecteur, nous précisons que dans cette section et la suivante, q désigne un ordre (expectile ou quantile) et non plus la fonction quantile comme dans la section précédente (où l'ordre était noté α).

2.2.2 Rappels sur la régression expectile

Soit $F(\cdot \mid \mathrm{X} = \mathrm{x})$ la fdr conditionnelle de Y sachant $\mathrm{X} = \mathrm{x}$ et considérons le problème de minimisation suivant :

$$\min_{\theta} \int \rho_q(y-\theta) dF(y \mid \mathrm{X} = \mathrm{x}) \tag{2.14}$$

où ρ_q est une fonction de perte avec q fixé, $0 < q < 1$. En dérivant par rapport à θ la fonction objectif dans l'équation (2.14), on obtient l'équation d'estimation suivante :

$$\int \psi_q(y-\theta) dF(y \mid \mathrm{X} = \mathrm{x}) = 0 \tag{2.15}$$

où $\psi_q(u) = \Delta\rho_q(u)/\Delta u$ est appelée la fonction d'influence. Si on définit $\psi_q(\cdot)$ égale à q pour les valeurs positives de son argument et à $-(1-q)$ pour les valeurs négatives, on obtient le quantile d'ordre q de la distribution conditionnelle $F(\cdot \mid \mathrm{X} = \mathrm{x})$ comme solution de (2.14) et (2.15). De manière similaire, l'expectile d'ordre q de cette distribution conditionnelle s'obtient comme solution des équations précédentes en posant :

$$\psi_q(u) = \begin{cases} qu & \text{si} \quad u \geq 0 \\ (1-q)u & \text{si} \quad u < 0 \end{cases} \tag{2.16}$$

dans (2.15). On peut remarquer que cela correspond à la fonction de perte des moindres carrés asymétriques

$$\rho_q(u) = \begin{cases} qu^2 & \text{si} \quad u \geq 0 \\ (1-q)u^2 & \text{si} \quad u < 0. \end{cases} \tag{2.17}$$

L'expectile conditionnel d'ordre q est unique (voir Newey et Powell 1987). Nous le noterons $m(q,\mathrm{x})$. L'expectile conditionnel d'ordre 0,5 est l'espérance de la distribution conditionnelle $F(\cdot \mid \mathrm{X} = \mathrm{x})$. Si l'on utilise dans (2.15) la fonction d'influence définie dans (2.16), on obtient alors une définition formelle de $m(q,\mathrm{x})$ comme solution de l'équation :

$$q = \frac{E(|Y - m(q,\mathrm{x})| \mathbb{1}(Y \leq m(q,\mathrm{x})) | \mathrm{X} = \mathrm{x})}{E(|Y - m(q,\mathrm{x})| \mid \mathrm{X} = \mathrm{x})}. \tag{2.18}$$

Newey et Powell 1987 ont montré que $m(q,\mathrm{x})$ est strictement croissante en q, ce qui permet d'utiliser les ordres-expectiles pour ordonner les observations (voir par exemple Kokic et al. 1997). Les propriétés théoriques des expectiles paramétriques peuvent être trouvées dans Newey et Powell 1987 et Efron 1991. Breckling et Chambers 1988 ont étendu le concept de la régression quantile et expectile aux M-quantiles et ont également défini un M-quantile dans le cas multivarié. Yao et Tong 1996 ont proposé un estimateur non paramétrique des expectiles conditionnels basé sur les polynômes locaux linéaires dans le cas d'une covariable unidimensionnelle et ont établi la normalité asymptotique et la convergence uniforme de leur estimateur.

2.2.3 Estimation des ordres-expectiles

Nous avons proposé cinq estimateurs différents de l'ordre-expectile d'une variable réponse Y univariée pour une covariable X dans $\mathbb{R}^p$. Les quatre premiers nécessitent l'estimation non paramétrique des expectiles conditionnels et comprennent donc deux étapes alors que le cinquième est obtenu par une approche directe.

Estimation des ordres-expectiles en deux étapes

La première étape consiste en l'estimation des expectiles conditionnels $m(q,\mathrm{x})$ sur une grille de valeurs de q. Ensuite, pour une valeur observée (y,x) quelconque, nous obtenons son ordre-expectile q par interpolation linéaire ou par une régression logistique.

Les quatre estimateurs se distinguent par la façon d'estimer les expectiles. Nous proposons pour la première étape :

1. un estimateur de Nadaraya-Watson localement constant,
2. un estimateur à noyaux localement linéaire.
   L'estimation par polynômes locaux ne donnant pas nécessairement un estimateur monotone par rapport à q, nous avons envisagé différentes techniques de monotonisation (voir Mukarjee et Stern 1994), qui conduisent aux deux estimateurs suivants :
3. un estimateur à noyaux localement linéaire monotone qui conserve la moyenne,
4. un estimateur à noyaux localement linéaire par régression isotonique.

Ces quatre techniques sont détaillées ci-dessous.

1. Régression expectile par la méthode de la constante locale

L'estimateur à noyaux $\widehat{m}_{\mathrm{LC}}(q,\mathrm{x})$ de $m(q,\mathrm{x})$ qui correspond à estimer cette fonction par une constante locale s'obtient comme la solution du problème de minimisation

$$\min_{\theta\in R}\int \rho_q(y-\theta)d\widehat{F}_n(y\mid \mathrm{X}=\mathrm{x}) \tag{2.19}$$

où

$$\widehat{F}_n(y\mid \mathrm{X}=\mathrm{x})=\frac{\sum_{i=1}^{n}K\left(\frac{\mathrm{x}-\mathrm{X}_i}{\mathrm{h}}\right)\mathbb{1}(Y_i\le y)}{\sum_{i=1}^{n}K\left(\frac{\mathrm{x}-\mathrm{X}_i}{\mathrm{h}}\right)}$$

est l'estimateur de Nadaraya-Watson de la fdr conditionnelle $F(\cdot\mid \mathrm{X}=\mathrm{x})$. K est un noyau multivarié et h est un vecteur de fenêtres adaptées. La fonction de perte ρ_q est celle définie par (2.17). En dérivant (2.19) par rapport à θ, on obtient l'équation d'estimation

$$\sum_{i=1}^{n}\psi_q(Y_i-\theta)K\left(\frac{\mathrm{x}-\mathrm{X}_i}{\mathrm{h}}\right)=0 \tag{2.20}$$

où ψ_q est définie dans (2.16). En posant $V_{q,i}(\mathrm{x})=(I_i-2qI_i+q)K\left(\frac{\mathrm{x}-\mathrm{X}_i}{\mathrm{h}}\right)$ où $I_i=\mathbb{1}(Y_i\le\theta)$ et en résolvant (2.20), on obtient l'estimateur $\widehat{m}_{\mathrm{LC}}(q,\mathrm{x})$ qui peut s'écrire comme une moyenne pondérée des observations de Y :

$$\widehat{m}_{\mathrm{LC}}(q,\mathrm{x})=\frac{\sum_{i=1}^{n}V_{q,i}(\mathrm{x})Y_i}{\sum_{i=1}^{n}V_{q,i}(\mathrm{x})}. \tag{2.21}$$

Pour un q quelconque, $\widehat{m}_{\mathrm{LC}}(q,\mathrm{x})$ s'obtient uniquement de manière itérative car $V_{q,i}(\mathrm{x})$ dépend de θ. On peut noter que cet estimateur est strictement croissant en q ce qui permet d'obtenir facilement l'estimateur $\widehat{q}_{\mathrm{LC}}(y,\mathrm{x})$ de l'ordre-expectile d'une observation y par interpolation linéaire sur une grille de valeurs de q définie pour chaque valeur de x.

2., 3. et 4. Ordres-expectiles basés sur des estimateurs localement linéaires

Une alternative à l'estimateur précédent est d'estimer la fonction expectile non paramétriquement en se basant sur un estimateur à noyaux localement linéaire (Yao et Tong 1996). Pour un vecteur u donné $p \times 1$, soit $\mathrm{u}^{*'} = [1\ \mathrm{u}']$. Un estimateur localement linéaire de $m(q,\mathrm{x})$ est donné par

$$\widehat{m}_{\mathrm{LL}}(q,\mathrm{x}) = \mathrm{x}^{*'}\widehat{\beta}_q(\mathrm{x}) \tag{2.22}$$

où $\widehat{\beta}_q(\mathrm{x})$ est la solution du problème de minimisation

$$\min_{\mathrm{b}\in R^{p+1}} \int \rho_q(y - \mathrm{x}^{*'}\mathrm{b}) d\widehat{F}_n(y \mid \mathrm{x}). \tag{2.23}$$

En dérivant (2.23) par rapport à b, on obtient l'équation d'estimation

$$\sum_{i=1}^{n} \psi_q(Y_i - \mathrm{X}_i^{*'}\mathrm{b}) K\left(\frac{\mathrm{x} - \mathrm{X}_i}{\mathrm{h}}\right) \mathrm{X}_i^{*'} = 0. \tag{2.24}$$

Soit Y le vecteur $n \times 1$ des observations de la variable réponse, $X^{*'} = [\mathrm{X}_1^{*'} \cdots \mathrm{X}_n^{*'}]$ avec les $\mathrm{X}_i^{*'}$ définis de manière identique à $\mathrm{u}^{*'}$ et soient $V_q(\mathrm{x})$ les matrices diagonales $n \times n$ des poids $V_{q,i}(\mathrm{x})$, où les $V_{q,i}(\mathrm{x})$ sont ceux de l'équation (2.21). La solution de (2.24) est alors

$$\widehat{\beta}_q(\mathrm{x}) = \left(X^{*'} V_q(\mathrm{x}) X^*\right)^{-1} X^{*'} V_q(\mathrm{x}) \mathrm{Y}.$$

Là encore $\widehat{\beta}_q(\mathrm{x})$ doit être calculé de façon itérative car la matrice $V_q(\mathrm{x})$ dépend de b.

Comme cet estimateur n'est pas nécessairement une fonction croissante de q, nous avons proposé deux méthodes de monotonisation. La première méthode consiste à contraindre l'estimateur $\widehat{m}_{\mathrm{LL}}(q,\mathrm{x})$ à être monotone pour les valeurs q d'une grille Q définie pour chaque valeur x de l'échantillon. Nous adaptons en fait une technique proposée par Mukarjee et Stern 1994 : pour tout q de la grille Q, l'estimateur $\widehat{m}_{\mathrm{LL}}(q,\mathrm{x})$ est remplacé par l'estimateur monotone $\widehat{m}_{\mathrm{MPM}}(q,\mathrm{x})$ qui conserve la moyenne :

$$\widehat{m}_{\mathrm{MPM}}(q,\mathrm{x}) = \begin{cases} \min_{q'\in Q,\ q\le q'\le 0{,}5} m_{\mathrm{LL}}(q',\mathrm{x}) & \text{si} \quad q \in]0\ ,\ 0{,}5], \\ \max_{q'\in Q,\ 0{,}5\le q'\le q} m_{\mathrm{LL}}(q',\mathrm{x}) & \text{si} \quad q \in]0{,}5\ ,\ 1[. \end{cases}$$

Une seconde approche pour construire un estimateur monotone de $m(q,x)$ est d'utiliser la régression isotonique (Robertson et al. 1988). Cela conduit à l'estimateur $\widehat{m}_{\mathrm{IRM}}(q,\mathrm{x})$ qui est l'estimateur monotone le plus proche de $m(q,x)$ au sens de la norme L_2. Soit $Q = \{q_1, \ldots, q_s\}$ une grille de valeurs de q avec $q_1 \le \cdots \le q_s$. Alors, pour q_i dans Q, $m_{\mathrm{IRM}}(q_i,x)$ est défini par

$$\widehat{m}_{\mathrm{IRM}}(q_i,\mathrm{x}) = \min_{\{i\le t\}} \max_{\{r\le i\}} Av\{\widehat{m}_{\mathrm{LL}}(q_k,\mathrm{x}),\ r \le k \le t\},$$

où $Av(X_1, \ldots, X_m)$ est la moyenne empirique de $X_1, \ldots, X_m$.

Quelle que soit la méthode de monotonisation employée, l'ordre-expectile de chaque observation (y,x) est ensuite obtenu par interpolation linéaire comme cela a été fait pour l'estimateur (2.21). Nous notons $\widehat{q}_{\mathrm{MPM}}(y,\mathrm{x})$ et $\widehat{q}_{\mathrm{IRM}}(y,\mathrm{x})$ les deux estimateurs de q obtenus.

Enfin, une alternative à la monotonisation est d'ajuster un modèle linéaire reliant les logits des valeurs de la grille Q aux valeurs estimées des expectiles conditionnels $\widehat{m}_{\mathrm{LL}}(q,\mathrm{x})$ calculées sur cette grille pour une valeur fixe des caractéristiques x. Nous obtenons alors l'ordre-expectile estimé $\widehat{q}_{\mathrm{LL}}(y,\mathrm{x})$ d'une observation (y,x) comme la valeur prédite par ce modèle pour la valeur y.

Estimation directe de l'ordre-expectile

D'après (2.18), la valeur y d'une observation (y,x) est l'expectile $m(q,\mathrm{x})$ où

$$q = \frac{E(|Y-y|\mathbb{1}(Y \leq y) \mid \mathrm{X}=\mathrm{x})}{E(|Y-y| \mid \mathrm{X}=\mathrm{x})}. \tag{2.25}$$

Pour chaque (y,x), nous estimons le numérateur et le dénominateur de (2.25) à l'aide d'estimateurs à noyaux de Nadaraya-Watson pondérés (Hall et al. 1999). Nous obtenons alors l'ordre-expectile par la formule suivante :

$$\widehat{q}_{\mathrm{ALNW}}(y,\mathrm{x}) = \frac{\sum_{i=1}^{n} |Y_i - y| \mathbb{1}(Y_i \leq y) K\left(\frac{\mathrm{x}-\mathrm{X}_i}{\mathrm{h}}\right) w_i(\mathrm{x})}{\sum_{i=1}^{n} |Y_i - y| K\left(\frac{\mathrm{x}-\mathrm{X}_i}{\mathrm{h}}\right) w_i(\mathrm{x})}, \tag{2.26}$$

où les $w_i(\mathrm{x})$ sont un ensemble de poids calibrés vérifiant $w_i \geq 0$, $\sum_i w_i = 1$ et

$$\sum_i (\mathrm{X}_i - \mathrm{x}) K\left(\frac{\mathrm{X}_i - \mathrm{x}}{\mathrm{h}}\right) w_i(\mathrm{x}) = 0. \tag{2.27}$$

La condition (2.27) assure que x est la moyenne pondérée des X_i avec les poids $\frac{K\left(\frac{\mathrm{X}_i-\mathrm{x}}{\mathrm{h}}\right) w_i(\mathrm{x})}{\sum_i K\left(\frac{\mathrm{X}_i-\mathrm{x}}{\mathrm{h}}\right) w_i(\mathrm{x})}$. Ces contraintes ne suffisent pas à définir ces poids de façon unique. Nous les avons donc calculés en minimisant la somme de leurs carrés. L'estimateur ainsi obtenu est une fonction croissante de y. De plus, comme le calcul de (2.26) est très rapide, on peut dériver par interpolation linéaire un estimateur $\widehat{m}_{\mathrm{ALNW}}(q,\mathrm{x})$ de $m(q,\mathrm{x})$ à partir de $\widehat{q}_{\mathrm{ALNW}}(y,\mathrm{x})$ calculé sur une grille très fine de valeurs de y, pour des valeurs fixées de q et de la covariable x.

2.2.4 Simulations

Les performances des cinq estimateurs des expectiles conditionnels ainsi que celles des estimateurs correspondants des ordres-expectiles ont été comparées à distance finie par simulation.

La comparaison des MASE montre que quel que soit q, l'estimateur localement linéaire $\widehat{m}_{\mathrm{LL}}$ a une meilleure performance que l'estimateur localement constant $\widehat{m}_{\mathrm{LC}}$ et que la monotonisation améliore le MASE, les estimateurs monotonisés $\widehat{m}_{\mathrm{MPM}}$ et $\widehat{m}_{\mathrm{IRM}}$ ayant des performances très proches. Quant à $\widehat{m}_{\mathrm{ALNW}}$, il se comporte mieux que les deux précédents pour des valeurs extrêmes de q mais est moins efficace pour les valeurs intermédiaires.

Pour les estimateurs des ordres-expectiles, ce sont les MADE (*mean absolute deviations errors*) qui ont été calculées pour chacun des 500 échantillons générés. D'après la figure 2.5, on peut constater que les estimateurs qui se comportent le mieux sont les estimateurs $\widehat{q}_{\mathrm{MPM}}$ and $\widehat{q}_{\mathrm{IRM}}$ basés sur la monotonisation des estimateurs localement linéaires des expectiles conditionnels.

2.2.5 Application

Les données fournies par l'URCAM concernent 2801 médecins généralistes de la région Midi-Pyrénées en 1999. La variable d'intérêt Y, qui mesure l'activité de prescription

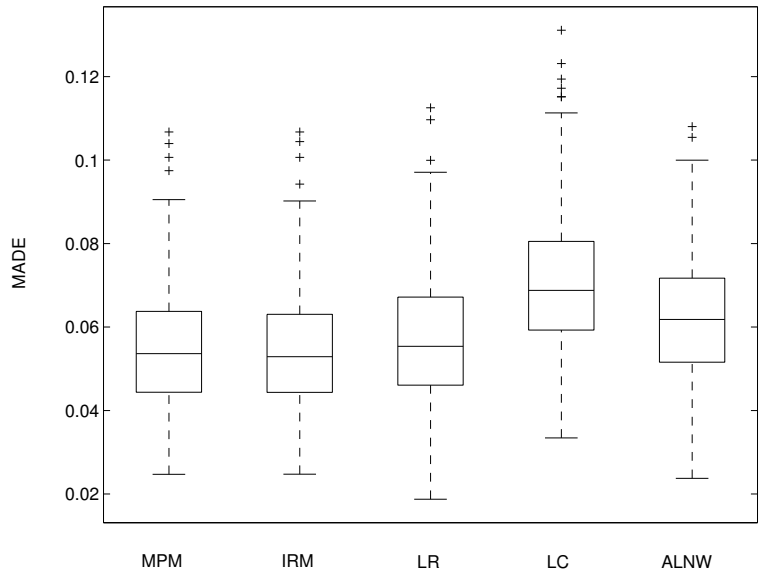


FIGURE 2.5 – Boîtes à moustaches des valeurs des MADE des estimateurs des ordres-expectiles pour les cinq méthodes proposées calculées sur les 500 échantillons générés. Les moyennes correspondantes valent respectivement 0,0545, 0,0544, 0,0575, 0,0708 et 0,0624.

pharmaceutique du médecin, est définie comme le logarithme du rapport du montant total des prescriptions pharmaceutiques de l'année sur le nombre total d'actes (consultations ou visites) du praticien pendant l'année. Nous disposons aussi de 15 variables décrivant l'activité et la clientèle du médecin ainsi que de son sexe et de son âge. Nous connaissons aussi le canton dans lequel le médecin exerce et avons des variables caractéristiques de ces cantons, ce qui nous donne deux niveaux de variables explicatives (au niveau du médecin et au niveau du canton).

Notre but est d'estimer pour chaque médecin son ordre-expectile conditionnellement à ses caractéristiques propres et de voir si la distribution de ces ordres diffère selon les cantons. Étant donné le nombre important de covariables qui peut rendre la régression expectile instable, nous avons dans une première étape réduit la dimension à l'aide de la régression inverse par tranches (méthode SIR, voir K.-C. LI 1991). Deux indices SIR ont été retenus pour résumer nos 15 covariables. Nous avons ensuite estimé les ordres-expectiles conditionnels à l'aide de l'estimateur $\widehat{q}_{\mathrm{MPM}}$. L'histogramme des ordres obtenus est montré à la figure 2.6. Les médecins dans les queues de la distribution montrent des comportements extrêmes en terme de prescription (en positif comme en négatif) par rapport aux autres médecins de Midi-Pyrénées qui partagent les mêmes caractéristiques. On peut remarquer que les médecins à « coût élevé » sont plus nombreux que ceux à « bas coût ». Cela n'aurait pas pu être montré avec des ordres-quantiles dont la distribution est nécessairement uniforme.

Pour voir si un effet canton peut expliquer les différences observées parmi les médecins, nous avons tracé les boîtes à moustaches des ordres-expectiles pour les 12 plus grands cantons de Midi-Pyrénées (voir figure 2.7). On constate des médianes assez contrastées : 0,8 pour le riche canton rural de Rodez contre 0,4 pour le canton de Auch, avec une médiane autour de 0,6 pour le canton de Toulouse. Une analyse de la variance du logit des ordres-expectiles en fonction du canton montre que les moyennes diffèrent

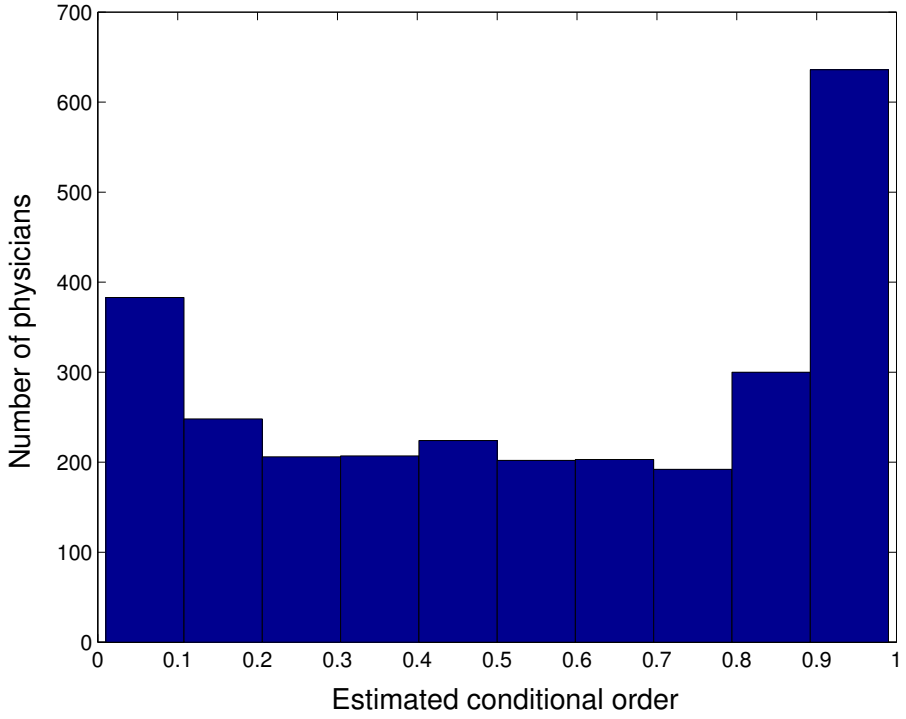


FIGURE 2.6 – Histogramme des ordres-expectiles estimés pour les 2801 médecins généralistes de Midi-Pyrénées.

significativement ($p = 0{,}03$).

Nous avons également estimé les ordres-expectiles en utilisant l'estimateur direct $\widehat{q}_{\text{ALNW}}$, dont le grand avantage est la rapidité de calcul (environ 1000 fois plus rapide que $\widehat{q}_{\text{MPM}}$). Le nuage de points de $\widehat{q}_{\text{ALNW}}$ en fonction de $\widehat{q}_{\text{MPM}}$ à la figure 2.8 permet de constater que ces deux estimateurs coïncident pour la plupart des médecins de notre échantillon, le coefficient de corrélation linéaire valant 0,99.

2.2.6 Conclusion et perspectives

L'approche que nous avons proposée permet de calculer non paramétriquement des ordres-expectiles et de les utiliser pour classer des individus. On peut ensuite étudier la liaison de ces mesures avec des variables géographiques ou contextuelles. Cette approche peut être vue comme une alternative non paramétrique intéressante aux méthodes paramétriques standard qui effectuent une modélisation à plusieurs niveaux. D'autre part, toutes les techniques utilisées peuvent se transposer en remplaçant les expectiles par des quantiles ou plus généralement des M-quantiles (BRECKLING et CHAMBERS 1988). L'intérêt de telles généralisations réside dans la robustesse aux valeurs extrêmes des estimateurs obtenus du fait que les M-quantiles sont basés sur des fonctions d'influence bornées.

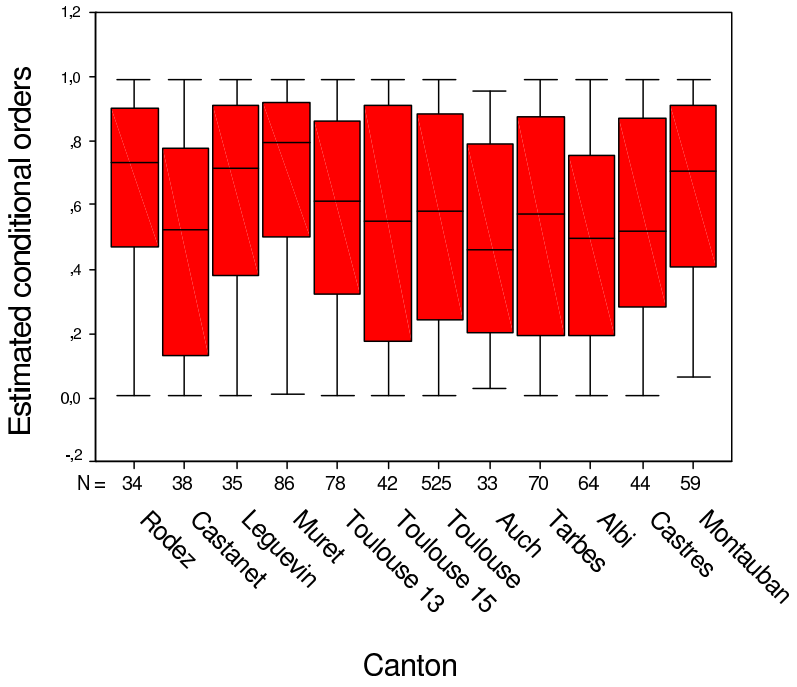


FIGURE 2.7 – Boîtes à moustaches des ordres-expectiles conditionnels estimés pour les 12 plus grands cantons de Midi-Pyrénées.

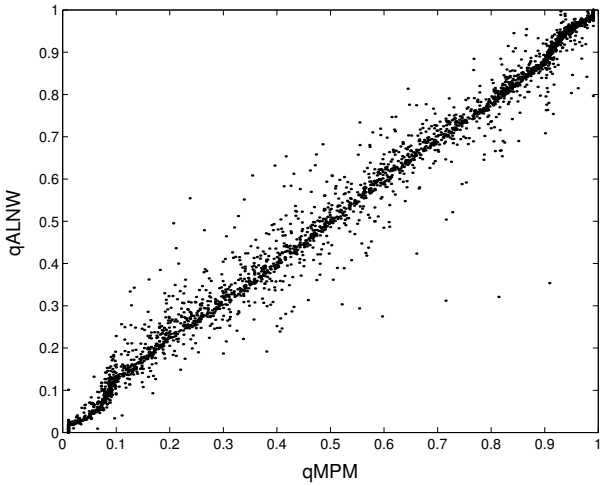


FIGURE 2.8 – Nuage de point de $\widehat{q}_{\text{ALNW}}$ en fonction de $\widehat{q}_{\text{MPM}}$.

2.3 Estimation de la fdr en sondages

2.3.1 Motivation : le cadre particulier des sondages

Les sondages constituent un cadre particulier : en effet, dans l'approche basée sur le plan de sondage, la population est considérée comme finie et on s'intéresse à une caractéristique des individus qui n'est pas considérée comme aléatoire. L'aléa vient uniquement de l'échantillon tiré. Une approche alternative est celle dite « basée sur un modèle », où on considère un modèle de superpopulation d'où est issue la population, la caractéristique d'intérêt étant alors considérée comme aléatoire. La littérature classique du domaine des sondages s'intéresse essentiellement à l'estimation de totaux ou de moyennes d'une caractéristique d'une population mais dans beaucoup d'applications, les paramètres d'intérêt peuvent être plus complexes : ce peut être des quantiles (voir par exemple Rueda et al. 2004) ou bien d'autres paramètres non linéaires dérivés de la fdr de la variable réponse. Avec ma collègue Sandrine Casanova, avec une approche basée sur un modèle, nous nous sommes intéressées à l'estimation de la fdr d'une population finie dans le cas où la variable réponse est censurée. Cela se produit quand la variable réponse correspond à une durée qui est observée pendant une période de temps limitée. Par exemple, si nous considérons des épisodes de chômage, les individus qui n'ont pas retrouvé d'emploi à la fin de leur période de suivi ont leur durée de chômage censurée à droite. Il faut bien distinguer le cas de la censure à droite de celui de la non-réponse : la censure donne accès à une information partielle alors qu'en cas de non réponse nous n'avons aucune information. A notre connaissance, il n'y a eu aucun article avant le nôtre traitant de l'estimation de la fdr d'une population finie en cas de censure à droite, sans doute parce que la méthodologie des données censurées s'est surtout développée dans le domaine médical où les sondages sont peu fréquents.

La section 2.3.2 présente l'article [10] qui propose un estimateur non paramétrique de la fdr d'une variable censurée en population finie avec une approche fondée sur un modèle. Le cas de l'estimation de la fdr sur petits domaines, qui correspond à la prépublication [22], est traité dans la section 2.3.3.

2.3.2 Estimation en population finie

L'estimation de la fdr dans le cadre des sondages a été abondamment étudiée en l'absence de censure. Pour une revue de ce qui a été proposé, on peut consulter le chapitre 6 de Mukhopadhyay 2001 ou le chapitre 36 de Pfefferman et C. R. Rao 2009. Une façon naïve d'estimer la fdr est de calculer la fdr empirique à partir des individus de l'échantillon. Dans l'approche basée sur le plan de sondage, l'estimateur habituel de la fdr est défini de la même façon que la fdr empirique mais prend en compte les probabilités d'inclusion du plan de sondage comme le fait l'estimateur de Horwitz-Thomson d'un total (Kuk 1988). Avec une approche intermédiaire, dite « assistée par un modèle », J. N. K. Rao et al. 1990 ont proposé un estimateur paramétrique de la fdr et une version non paramétrique de cet estimateur a été définie par Johnson et al. 2008. Dans toute la suite, nous allons nous restreindre à l'approche basée sur un modèle. Chambers et Dunstan 1986 montrent qu'on améliore l'estimation de la fdr en prédisant les valeurs de la variable réponse pour les individus non échantillonnés à l'aide d'une régression paramétrique qui utilise une information auxiliaire. Cet estimateur sera noté dans la suite avec l'indice cd. S. Wang et Dorfman 1996 ont proposé d'utiliser une moyenne pondérée de l'estimateur cd et de celui de J. N. K. Rao et al. 1990, ce qui permet d'améliorer l'erreur quadratique moyenne. Plusieurs variantes de l'estimateur cd et de l'estimateur de J. N. K. Rao et al. 1990 ont été proposées (voir le chapitre 36 de Pfefferman et C. R. Rao 2009). Dorfman et Hall 1993 ont défini une version non

paramétrique de l'estimateur CD et en ont étudié les propriétés asymptotiques.

Dans [10], nous avons proposé un estimateur non paramétrique de la fdr pour une variable réponse censurée à droite avec l'approche basée sur un modèle. L'estimateur utilise de l'information auxiliaire fournie par une covariable continue et se base sur la régression médiane non paramétrique adaptée au cas censuré. Après avoir défini les notations et rappelé l'estimation naïve que nous pouvons faire de la fdr, nous détaillons ci-dessous le principe de cet estimateur et proposons une estimation bootstrap du biais et de la variance de l'erreur de prédiction. Un exemple d'application et des simulations sont ensuite présentés.

Notations

Nous considérons une population finie $\mathcal{P}$ de taille N dans laquelle est tiré un échantillon s de taille n. t_j est la valeur de la variable d'intérêt pour l'individu j de $\mathcal{P}$. Nous supposons que t_j est positif et peut être censuré à droite par un délai de censure z_j. Ainsi, pour $j \in s$, nous observons $y_j = \min(t_j, z_j)$ et $\delta_j = \mathbb{1}(t_j \leq z_j)$. En outre, nous supposons que nous disposons d'une information auxiliaire, connue pour toute la population, sous la forme d'une variable X continue dont la valeur sera notée x_j pour l'individu j.

En population finie, la fdr de la variable réponse T que nous cherchons à estimer est $F(t) = \dfrac{1}{N} \sum_{j \in \mathcal{P}} \mathbb{1}(t_j \leq t)$.

Un estimateur naïf de la fdr

Il est bien connu que la fonction de répartition empirique ne fournit pas un estimateur convergent de la fdr en présence de censure. Par contre, le complémentaire à 1 de l'estimateur de Kaplan-Meier (KAPLAN et MEIER 1958) de la fonction de survie calculé sur l'échantillon s constitue un estimateur naïf convergent de F.

Comme l'estimateur original de Kaplan-Meier 1.1 n'est pas défini après le dernier délai observé $y_{(n)}$ si celui-ci correspond à une censure, afin d'obtenir une fonction de répartition, nous allons lui substituer la version d'EFRON 1967 définie par :

$$\widehat{F}_{\text{KM}}(t) = \begin{cases} 1 - \displaystyle\prod_{j \in s} \left\{ 1 - \frac{1}{\displaystyle\sum_{r \in s} \mathbb{1}(y_r \geq y_j)} \right\}^{\mathbb{1}(y_j \leq t, \delta_j = 1)} & \text{si } \; t < y_{(n)}, \\ 1 & \text{sinon.} \end{cases} \tag{2.28}$$

Le nouvel estimateur

L'estimateur que nous proposons utilise l'approche basée sur un modèle. Nous n'utiliserons donc pas les poids de sondage et n'avons donc pas besoin de nous placer dans un contexte de plan de sondage particulier. Cependant, pour obtenir un estimateur convergent et efficace, nous devons supposer que le plan de sondage est non informatif, ce qui signifie que le modèle sur lequel nous nous basons est valide à la fois pour l'échantillon et la population (voir l'introduction de la partie 4 de PFEFFERMAN et C. R. RAO 2009). Cela justifie aussi le choix d'un estimateur non paramétrique pour lequel le risque de mauvaise spécification est réduit.

La fdr F peut se décomposer en deux termes :

$$F(t) = \frac{1}{N} \sum_{j \in s} \mathbb{1}(t_j \leq t) + \frac{1}{N} \sum_{j \in \mathcal{P} \setminus s} \mathbb{1}(t_j \leq t). \tag{2.29}$$

Le premier terme n'est pas connu à cause de la censure à droite et doit donc être estimé. Comme il peut s'écrire :

$$\frac{1}{N}\sum_{j\in s}\mathbb{1}(t_j \leq t) = \frac{n}{N}\left(\frac{1}{n}\sum_{j\in s}\mathbb{1}(t_j \leq t)\right), \tag{2.30}$$

nous reconnaissons la fdr basée sur l'échantillon s entre les parenthèses, qui peut donc être estimée par l'estimateur de Kaplan-Meier calculé sur les individus de l'échantillon s.

Pour pouvoir estimer non paramétriquement le second terme de (2.29), nous postulons le modèle de superpopulation suivant :

$$\xi : \quad t_j = m(x_j) + \varepsilon_j \quad (j \in \mathcal{P}) \tag{2.31}$$

où les ε_j sont des variables i.i.d. de fdr G et où $m(x_j)$ est la médiane conditionnelle de T sachant $X = x_j$. Le choix de modéliser la relation entre T et X par la médiane conditionnelle plutôt que par la moyenne conditionnelle est dû au fait qu'en présence de données censurées, la médiane est un paramètre de position plus utilisé et plus robuste que la moyenne.

Comme $E(\mathbb{1}(t_j \leq t)) = P(t_j \leq t) = G(t - m(x_j))$, une prédiction de $\mathbb{1}(t_j \leq t)$ peut être obtenue en estimant $G(t - m(x_j))$. Nous devons dans un premier temps estimer la médiane conditionnelle $m(x_j)$. Dans ce but, nous estimons la fdr conditionnelle de T sachant $X = x$ à l'aide de l'estimateur généralisé de Kaplan-Meier (voir formule 2.1), calculé sur l'échantillon s. L'estimateur est donc $\widehat{F}_{\text{GKM}}(t \mid x) = 1 - \widehat{S}_{\text{GKM}}(t \mid x)$, où nous choisissons pour les $B_j(x)$ des poids de type Nadaraya-Watson (voir 2.2).

Pour pouvoir estimer la médiane conditionnelle par inversion, nous utilisons la version lissée (2.3) de l'estimateur $\widehat{F}_{\text{GKM}}$ que nous avons proposée dans l'article [3] et que nous notons $\widehat{F}_{\text{SGKM}}$. Nous définissons donc l'estimateur de la médiane conditionnelle par $\widehat{m}(x_j) = \widehat{F}^{-1}_{\text{SGKM}}(0{,}5 \mid x_j)$.

Revenons à l'estimation de $G(t - m(x_j))$. Comme les résidus $\widehat{\varepsilon}_j = y_j - \widehat{m}(x_j)$, $j \in s$, sont censurés à droite si les y_j le sont, un estimateur convergent de la fdr G des erreurs est l'estimateur de Kaplan-Meier calculé à partir des résidus $\widehat{\varepsilon}_j$, noté $\widehat{G}_{\text{KM}}$. Nous en déduisons alors un estimateur de F :

$$\widehat{F}_{\text{M}}(t) = \frac{1}{N}\left(n\widehat{F}_{\text{KM}}(t) + \sum_{j\in\mathcal{P}\setminus s}\widehat{G}_{\text{KM}}(t - \widehat{m}(x_j))\right). \tag{2.32}$$

$\widehat{F}_{\text{M}}$ est une fonction croissante qui converge vers 1 quand t tend vers l'infini. Nous obtenons donc bien une fdr.

Nous montrons dans [3] comment cet estimateur peut être adapté très simplement au cas d'un plan de sondage stratifié ainsi qu'à la prise en compte de la non-réponse complète ou partielle.

Comportement asymptotique

A cause de la censure qui vient compliquer le calcul de l'estimateur, nous n'avons pas pu obtenir de résultats asymptotiques concernant le biais sous le modèle de l'estimateur $\widehat{F}_{\text{M}}$. Nous pouvons cependant noter que cet estimateur est très proche de la version non paramétrique de l'estimateur CD proposée par Dorfman et Hall 1993, qui ont montré que cet estimateur était asymptotiquement sans biais sous le modèle sous certaines conditions concernant les fenêtres. Ils ont également établi un développement asymptotique pour la variance de leur estimateur, ce qui permet d'en montrer la convergence.

Estimation bootstrap du biais et de la variance de l'erreur de prédiction

Dans le cadre des sondages, on définit l'erreur de prédiction d'un estimateur $\widehat{F}$ de la fdr F par $\widehat{F}(t) - F(t)$ et l'on cherche à estimer le biais et la variance de l'erreur de prédiction. En raison de la complexité de nos estimateurs, nous n'avons pas pu obtenir de formules explicites pour estimer ces quantités. Cependant, Lombardia et al. 2004 ont proposé d'utiliser des techniques bootstrap pour estimer le biais et la variance de l'erreur de prédiction de l'estimateur non paramétrique cd. Nous avons adapté leur méthodologie au cas censuré et obtenu des estimations bootstrap du biais et de la variance de l'erreur de prédiction de $\widehat{F}_{\mathrm{M}}$. L'idée clé vient d'un argument proposé par Booth et al. 1994 : pour estimer une caractéristique d'une population finie, on peut utiliser la moyenne des valeurs de cette caractéristique calculée sur des populations échantillonnées à partir de l'échantillon initial. Cette technique nous a en outre permis d'obtenir un intervalle de confiance bootstrap pour F.

Exemple d'application

Nous avons considéré des données de panels du sipp (Survey of Income and Program Participation) de Hu et Ridder 2012. Nous nous sommes limitées aux panels de 1992 et 1993 et à l'échantillon des familles monoparentales qui ont bénéficié du programme d'aide sociale pour les familles avec enfants à charge. La variable d'intérêt est la durée du premier épisode d'aide financière pour la famille. Chaque famille étant suivie pendant 36 mois, cette durée est censurée si la famille quitte le panel avant la fin de son allocation. Nous avons dans ce sous-échantillon 520 épisodes d'aide dont 269 sont censurés ce qui conduit à un taux de censure de 51,7 %. La variable auxiliaire que nous avons choisie est le montant de l'aide, variable qui est significativement et négativement liée à la probabilité de quitter le programme d'aide sociale (voir Fitzgerald 1991). Comme cette variable doit être connue sur toute la population, nous avons considéré comme population fixée $\mathcal{P}$ notre ensemble de 520 épisodes d'aide. Nous présentons ici les résultats pour un plan de sondage aléatoire simple : 40 individus sont tirés aléatoirement sans remise dans $\mathcal{P}$. Nous avons calculé les estimateurs $\widehat{F}_{\mathrm{KM}}$ et $\widehat{F}_{\mathrm{M}}$ en choisissant les fenêtres par validation croisée. Comme nous ne connaissons pas la vraie fdr F étant donné que les valeurs de T dans notre population sont censurées à droite, nous utilisons l'estimateur de Kaplan-Meier $\widehat{F}_{\mathrm{N}}$ de la fdr calculé sur tout $\mathcal{P}$ comme fdr cible. La figure 2.9 présente les graphes des deux estimateurs de la fdr ainsi que les intervalles de confiance (IC) bootstrap à 95 % de F basés sur ces deux estimateurs. On peut constater que l'IC basé sur $\widehat{F}_{\mathrm{M}}$ est toujours plus étroit que celui basé sur $\widehat{F}_{\mathrm{KM}}$. On peut par inversion des estimateurs des fdr obtenir des estimations des durées médianes des épisodes d'aide. La médiane basée sur $\widehat{F}_{\mathrm{KM}}$ vaut 6,68 mois contre 10,88 mois pour celle basée sur $\widehat{F}_{\mathrm{M}}$. Cette dernière estimation est très proche de la médiane basée sur notre cible $\widehat{F}_{\mathrm{N}}$ qui vaut 10,79 mois.

Simulations

Dans le cadre des sondages, on peut effectuer des simulations basées sur le modèle, où la population est générée à chaque itération, ou bien des simulations basées sur le plan de sondage, où tous les échantillons sont tirés d'une unique population. Nous avons considéré les deux approches.

Les simulations basées sur le modèle ont permis de faire varier différents paramètres comme le taux de censure, la taille de la population et la force de la relation entre la variable réponse et la variable auxiliaire. Nous avons comparé les performances de l'estimateur $\widehat{F}_{\mathrm{M}}$ et de l'estimateur naïf de Kaplan-Meier calculé à partir des points échantillonnés. Le nouvel estimateur est biaisé mais il présente une plus faible erreur quadratique moyenne que l'estimateur de Kaplan-Meier, l'écart augmentant avec la force

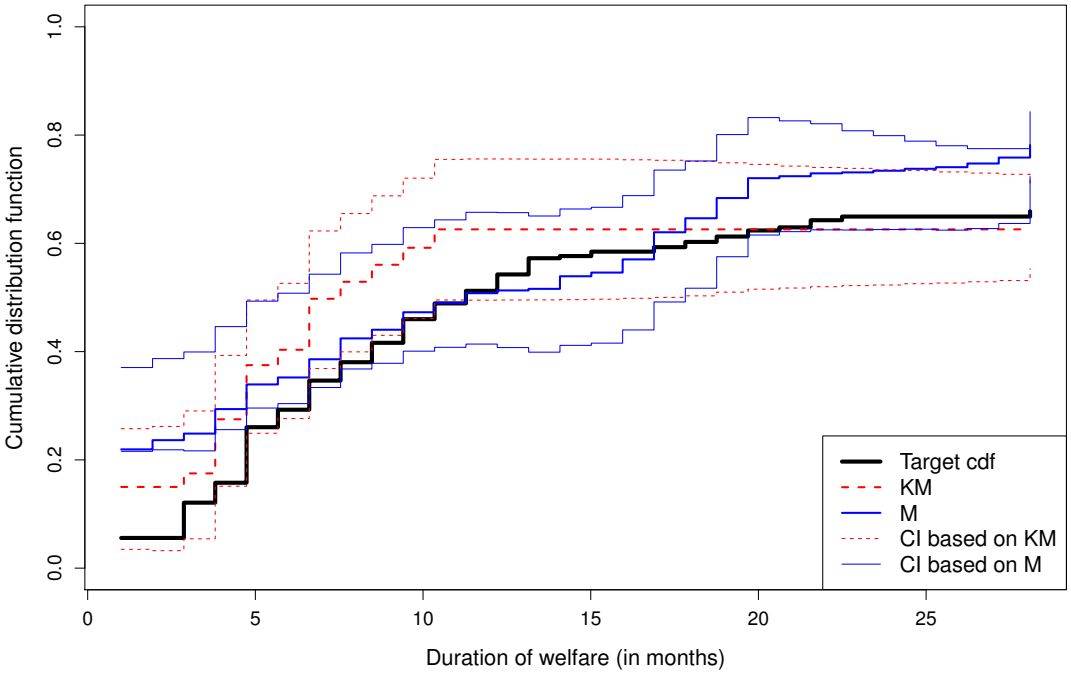


FIGURE 2.9 – Estimateurs de la fdr de la durée de l'épisode d'aide financière et les IC bootstrap correspondants à 95 % basés sur un échantillon de taille $n = 40$ pour un plan de sondage aléatoire simple. La fdr cible est représentée par l'estimateur de Kaplan-Meier $\widehat{F}_{\mathrm{N}}$ calculé sur les 520 individus de l'échantillon SIPP.

de la relation entre la variable réponse et la variable auxiliaire.

Les simulations basées sur le plan de sondage ont utilisé l'échantillon des 520 épisodes d'aide du SIPP présenté dans l'exemple ci-dessus, considéré comme notre population, dans laquelle nous tirons 500 échantillons selon deux plans de sondage : un plan de sondage aléatoire simple et un plan de sondage stratifié avec deux strates dépendant du montant de l'aide et allocation proportionnelle. Le rapport des MASE de l'estimateur $\widehat{F}_{\mathrm{KM}}$ et de l'estimateur $\widehat{F}_{\mathrm{M}}$ vaut 1,87 pour le plan simple et 1,63 pour le plan stratifié, montrant le gain obtenu avec l'estimateur $\widehat{F}_{\mathrm{M}}$.

2.3.3 Estimation sur petits domaines

Un axe de recherche important en sondages est l'estimation sur petits domaines, qui correspond à l'estimation des quantités d'intérêt sur des sous-populations de petite taille. Si un domaine est de taille suffisante, l'estimation des paramètres d'intérêt peut se restreindre aux données relatives aux individus du domaine et les estimateurs produits sont de précision acceptable. Cependant, dans la plupart des applications, les tailles d'échantillons correspondant à des petits domaines ne sont pas suffisantes. L'estimation se fonde alors sur une information auxiliaire fournie par une covariable et de l'information est « empruntée » aux autres domaines. L'ensemble de ces techniques est appelé « estimation sur petits domaines ».

Le modèle classique qui permet de capturer l'effet domaine est le modèle linéaire

mixte (voir J. N. K. Rao 2003). L'estimation des coefficients de régression et la prédiction des effets aléatoires s'obtiennent par maximum de vraisemblance sous l'hypothèse de normalité des erreurs, conduisant au meilleur prédicteur empirique linéaire sans biais (EBLUP) de la variable d'intérêt. Cependant, ces modèles reposent sur des hypothèses fortes comme la normalité et l'homoscédasticité des erreurs. Pour prendre en compte une relation non-linéaire entre la variable d'intérêt et la variable auxiliaire, Salvati, Chandra, Ranalli et al. 2010 ont proposé une version non paramétrique de l'EBLUP pour la moyenne sur un petit domaine en utilisant des splines pénalisés. Dans un cadre non paramétrique, Salvati, Ranalli et al. 2010 estiment la moyenne sur un petit domaine par des M-quantiles conditionnels basés sur des splines pénalisés. En ce qui concerne l'estimation de la fdr, Chambers et Tzavidis 2006 prédisent la fdr d'un petit domaine à l'aide de M-quantiles conditionnels paramétriques. Casanova 2012 a étendu la technique de Chambers et Tzavidis 2006 au cas non paramétrique en utilisant des estimateurs à noyaux des M-quantiles conditionnels. Enfin, Salvati, Chandra et Chambers 2010 proposent d'estimer la fdr sur un petit domaine par la moyenne des données de l'échantillon du petit domaine pondérée par des poids d'échantillonnage calibrés.

Nous proposons dans [22] un estimateur non paramétrique de la fdr sur petits domaines dans le cas où la variable d'intérêt est censurée à droite, cas qui, à notre connaissance, n'a jamais été considéré dans la littérature. Après avoir défini les notations dans ce nouveau cadre, nous détaillons la construction de l'estimateur, qui est une adaptation au cas censuré de la technique de Casanova 2012. Un exemple d'application à des données sur les temps d'accès au premier emploi de jeunes diplômées illustre la méthode. Enfin, des simulations basées sur le modèle comparent cet estimateur avec l'estimateur naïf de Kaplan-Meier calculé sur les points échantillonnés du domaine et l'estimateur de la section précédente $\widehat{F}_{\mathrm{M}}$ calculé sur le domaine. Nous finissons la section par quelques perspectives.

Notations

La population $\mathcal{P}$ de la section précédente est maintenant partitionnée en m sous-populations — appelées domaines — U_i de taille N_i, $i = 1, \ldots, m$. Soient s un échantillon de $\mathcal{P}$ de taille n et $s_i = s \cap U_i$ un échantillon du domaine U_i de taille n_i. t_{ij} est la variable d'intérêt mesurée pour le j-ème individu du domaine U_i. On suppose que t_{ij} est seulement connu sur s_i et éventuellement censuré à droite par z_{ij}. Avec les notations d'Efron, nous observons, sur l'échantillon s_i, $y_{ij} = \min(t_{ij}, z_{ij})$ et $\delta_{ij} = \mathbb{1}(t_{ij} < z_{ij})$. Nous disposons en outre d'une information auxiliaire x_{ij}, valeur d'une variable continue X pour l'individu j du domaine i, connue sur toute la population.

Dans le cadre des sondages, la fdr de la variable d'intérêt T sur le domaine U_i s'écrit $F^i(t) = \dfrac{1}{N_i} \displaystyle\sum_{j \in U_i} \mathbb{1}(t_{ij} \leq t)$ que l'on peut décomposer en

$$F^i(t) = \frac{1}{N_i} \left(\sum_{j \in s_i} \mathbb{1}(t_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} \mathbb{1}(t_{ij} \leq t) \right). \tag{2.33}$$

Le nouvel estimateur

Comme dans la section précédente, nous pouvons estimer F^i par l'estimateur de Kaplan-Meier $\widehat{F}^i_{\mathrm{KM}}$ calculé à partir des points de l'échantillon s_i du domaine U_i. On peut sans doute améliorer l'estimation de la fdr F^i en construisant des estimateurs basés sur un modèle, qui utilisent de l'information auxiliaire, en se basant sur la formule (2.33).

Un estimateur naturel de F^i peut être obtenu en remplaçant s par s_i dans toutes les formules de la section 2.3.2. L'estimateur qui en résulte sera noté $\widehat{F}^i_{\mathrm{M}}$. Les fenêtres h_T et h_X doivent être remplacées par des fenêtres adaptées à chaque domaine, notées h^i_T et h^i_X. Cette approche nécessite de postuler le modèle de superpopulation ξ' suivant :

$$t_{ij} = m(x_{ij}) + e_{ij},\ i = 1,\ldots,m,\ \ j = 1,\ldots,N_i,$$

où les e_{ij} sont des variables i.i.d. de fdr G'^i et où $m(x_{ij})$ désigne la médiane conditionnelle de T sachant $X = x_{ij}$.

Cependant, quand la taille du domaine est petite, ces estimateurs peuvent avoir une grande variance et des méthodes qui utilisent l'information apportée par les autres domaines doivent être préférées pour améliorer la précision de l'estimation. Pour ce faire, nous proposons la procédure suivante. D'après l'équation (2.30) appliquée au domaine U_i, le premier terme de (2.33) peut toujours être estimé par l'estimateur de Kaplan-Meier calculé sur l'échantillon s_i. En ce qui concerne le second terme de (2.33), nous utilisons l'information de l'échantillon total s pour prédire la valeur de la variable réponse pour les individus non échantillonnés du domaine U_i. Dans ce cadre, nous devons supposer le modèle de superpopulation ζ :

$$t_{ij} = m(q_i,x_{ij}) + \varepsilon_{ij},\quad i = 1,\ldots,m,\quad j = 1,\ldots,N_i,$$

où les ε_{ij} sont des variables i.i.d. de fdr G^i ; q_i est un coefficient de [0,1] caractérisant la position du domaine U_i et $m(q_i,x_{ij})$ est le quantile conditionnel d'ordre q_i de T sachant $X = x_{ij}$.

Chaque valeur t_{ij} peut en effet être considérée comme le quantile conditionnel de T sachant $X = x_{ij}$ pour un certain ordre noté $q(t_{ij},x_{ij})$. De ce fait, en imitant Chambers et Tzavidis 2006, le coefficient q_i du domaine U_i peut être défini comme la moyenne ou la médiane des ordres-quantiles conditionnels $q(t_{ij},x_{ij})$ des individus j du domaine U_i. On peut noter que les ordres-quantiles conditionnels sont définis au niveau de la population et que nous nous attendons à ce que des individus du même domaine aient des valeurs de leurs ordres-quantiles assez proches si une part de la variabilité des données est expliquée par le domaine.

Les ordres-quantiles conditionnels sont estimés à l'aide de la version lissée de l'estimateur de Kaplan-Meier généralisé calculé sur l'échantillon total s selon la formule (2.3) :

$$\widehat{q(t_{ij},x_{ij})} = \widehat{F}_{\mathrm{SGKM}}(y_{ij} \mid x_{ij}).$$

Comme les valeurs y_{ij} peuvent être censurées à droite, les ordres $\widehat{q(t_{ij},x_{ij})}$ le sont aussi. Ainsi, pour estimer l'ordre global q_i du domaine U_i par la moyenne ou la médiane des ordres $\widehat{q(t_{ij},x_{ij})}$, nous devons d'abord estimer leur fdr en tenant compte des observations censurées. Cela peut être facilement réalisé en utilisant une fois encore l'estimateur de Kaplan-Meier. Comme la médiane est plus facile à estimer que la moyenne en présence de censure, nous choisissons donc d'estimer q_i par la médiane $\widehat{q}_i$ des ordres-quantiles, obtenue en inversant l'estimateur de Kaplan-Meier.

Comme $E_\zeta\left(\mathbb{1}(t_{ij} \le t)\right) = P(t_{ij} \le t) = G^i(t - m(q_i,x_{ij}))$, $\mathbb{1}(t_{ij} \le t)$ peut être prédite en estimant $G_i(t-m(q_i,x_{ij}))$. Un estimateur naturel $\widehat{m}(\widehat{q}_i,x_{ij})$ de $m(q_i,x_{ij})$ est le quantile conditionnel d'ordre $\widehat{q}_i$ sachant x_{ij}, qui est la solution en θ de $\widehat{F}_{\mathrm{SGKM}}\left(\theta \mid x_{ij}\right) = \widehat{q}_i$ et est donc obtenue en inversant $\widehat{F}_{\mathrm{SGKM}}$. On peut remarquer que, là encore, comme pour l'estimation de l'ordre-quantile q_i, l'échantillon tout entier est utilisé pour calculer cet estimateur, ce qui permet d'emprunter de la force aux autres domaines. Comme dans la

section 2.3.2, $G^i(t-m(q_i,x_{ij}))$ peut être estimé par l'estimateur de Kaplan-Meier calculé à partir des résidus censurés $\widehat{\varepsilon}_{ij} = y_{ij} - \widehat{m}(\widehat{q}_i,x_{ij})$, $j \in s_i$. Nous notons cet estimateur $\widehat{G}^i_{\mathrm{KM}}$ et en déduisons l'estimateur suivant de la fdr de T dans le domaine U_i :

$$\widehat{F}^i_{\mathrm{Q}}(t) = \frac{1}{N_i}\left(n_i \widehat{F}^i_{\mathrm{KM}}(t) + \sum_{j\in U_i\setminus s_i} \widehat{G}^i_{\mathrm{KM}}\left(t - \widehat{m}(\widehat{q}_i,x_{ij})\right)\right). \tag{2.34}$$

Cet estimateur est une fdr de façon évidente.

Application

Nous avons analysé par nos méthodes des données du Centre d'études et de recherches sur les qualifications (Céreq), en collaboration avec Hélène Couprie, enseignant-chercheur en économie. Le Céreq interroge par sondage les jeunes diplômés sur leur devenir professionnel trois ans après l'obtention de leur diplôme (étude rétrospective sur la situation mensuelle des trois années précédentes). Nous nous sommes restreintes aux 10135 jeunes filles des régions Midi-Pyrénées et Languedoc-Roussillon qui sont sorties de l'enseignement secondaire en 2010. La variable d'intérêt T est le temps d'accès au premier emploi depuis l'obtention du diplôme, censuré à droite pour les jeunes filles qui n'ont pas trouvé d'emploi à la fin de l'enquête (12,5 % dans nos données). Le Céreq est intéressé par avoir des statistiques selon le niveau et le type de formation suivi, variable qui nous a partitionné la population en 34 domaines (de taille variant de 7 à 1480 jeunes filles). L'échantillon, d'un effectif total de 306, se partitionne par domaine en sous-échantillons de taille 1 à 37. La variable auxiliaire choisie est le taux de chômage local de la zone d'emploi de l'établissement de fin d'études, qui est significativement et négativement lié à la probabilité de trouver un emploi (p=0,014). La figure 2.10 présente les fonctions de survie correspondant aux trois estimateurs $\widehat{F}^i_{\mathrm{KM}}$, $\widehat{F}^i_{\mathrm{M}}$ et $\widehat{F}^i_{\mathrm{Q}}$ pour 6 domaines de tailles très différentes.

Simulations

Des simulations basées sur le modèle sont en cours, en simulant un modèle log-linéaire de régression avec un effet aléatoire du domaine. Les paramètres qui varient sont le taux de censure, la taille des échantillons des domaines, la force de la relation entre la variable d'intérêt et la variable auxiliaire mais aussi la part de la variabilité due aux domaines. Les premiers résultats montrent que l'estimateur $\widehat{F}^i_{\mathrm{M}}$ est toujours meilleur que l'estimateur naïf de Kaplan-Meier sur le domaine. Le nouvel estimateur $\widehat{F}^i_{\mathrm{Q}}$ qui emprunte de la force aux voisins est plus performant que $\widehat{F}^i_{\mathrm{M}}$ quand l'échantillon du domaine est de très petite taille (inférieure à 6 individus) et que la variabilité due aux domaines est faible ou modérée. Quand les domaines diffèrent plus fortement, les deux estimateurs sont équivalents.

2.3.4 Conclusion et perspectives

Comme nous l'avons fait dans [10] en population globale, des techniques de type bootstrap pourraient être adaptées au cas des petits domaines pour obtenir des intervalles de confiance de la fdr du domaine, ainsi que des estimations du biais et de la variance de l'erreur de prédiction.

Une approche intermédiaire qui utilise les poids de sondage tout en se basant sur un modèle est l'approche dite « assistée par un modèle ». Il serait intéressant d'adapter les estimateurs proposés dans ce contexte au cas censuré et de les comparer aux estimateurs basés sur un modèle proposés dans [10] et [22].

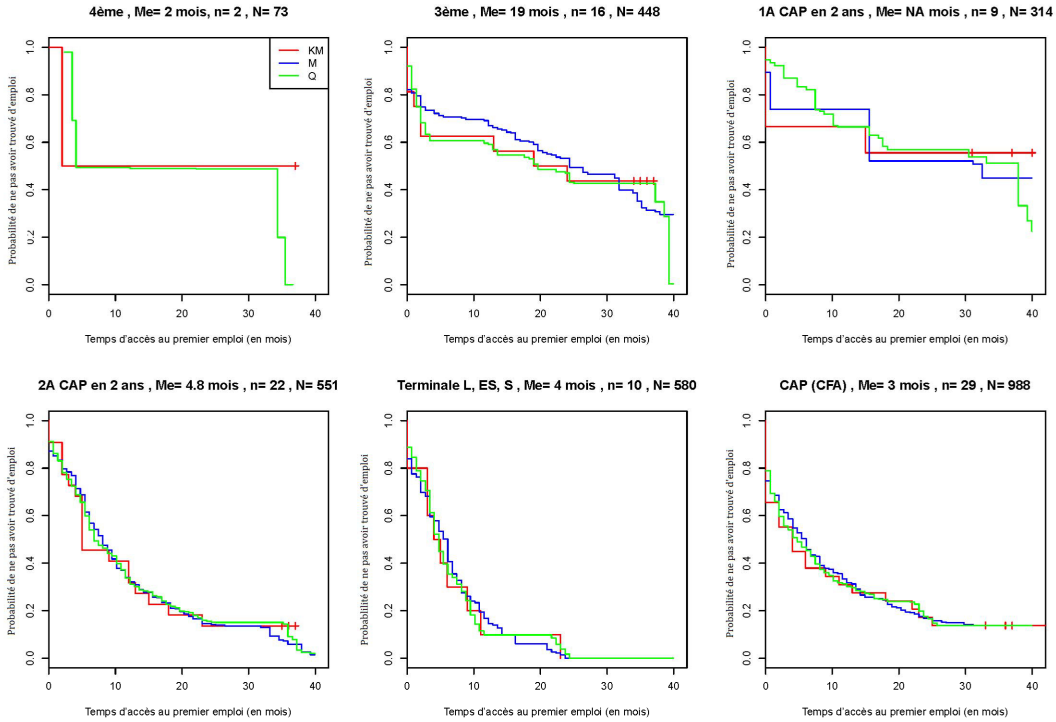


FIGURE 2.10 – Probabilité de ne pas avoir trouvé de premier emploi, estimée par les trois estimateurs pour six domaines différents correspondant à un niveau et type de formation. La médiane indiquée (Me) est obtenue en inversant l'estimateur naïf de Kaplan-Meier.

2.4 Références

BERAN, R. (1981). *Nonparametric regression with randomly censored survival data*. Rapp. tech. University of California, Berkeley.

BERLINET, A., A. GANNOUN et E. MATZNER-LØBER (1998). « Propriétés asymptotiques d'estimateurs convergents des quantiles conditionnels ». In : *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 326.5, p. 611–614.

BOOTH, J. G., R. W. BUTLER et P. HALL (1994). « Bootstrap Methods for Finite Populations ». In : *Journal of the American Statistical Association* 89.428, p. 1282–1289. URL : http://www.jstor.org/stable/2290991.

BRECKLING, J. et R. CHAMBERS (1988). « M-quantiles ». In : *Biometrika* 75.4, p. 761–771. URL : http://dx.doi.org/10.1093/biomet/75.4.761.

CAMPBELL, G. (1981). « Nonparametric Bivariate Estimation with Randomly Censored Data ». In : *Biometrika* 68.2, p. 417–422. URL : http://www.jstor.org/stable/2335587.

CAMPBELL, G. et A. FÖLDES (1982). « Large-sample properties of nonparametric bivariate estimators with censored data ». In : *Nonparametric Statistical Inference, Colloquia Mathematica- Societatis János Bolyai*, p. 103–122.

CASANOVA, S. (2012). « Using Nonparametric Conditional M-Quantiles to Estimate a Cumulative Distribution Function in a Domain ». In : *Annals of Economics and Statistics* 107/108, p. 287–297. URL : http://www.jstor.org/stable/23646580.

Chambers, R. et R. Dunstan (1986). « Estimating distribution functions from survey data ». In : *Biometrika* 73.3, p. 597–604. doi : 10.1093/biomet/73.3.597.

Chambers, R. et N. Tzavidis (2006). « M-quantile models for small area estimation ». In : *Biometrika* 93.2, p. 255–268. doi : 10.1093/biomet/93.2.255.

Dabrowska, D. M. (1989). « Uniform consistency of the kernel conditional Kaplan-Meier estimate ». In : *Annals of Statistics* 17, p. 1157–1167.

— (1992). « Nonparametric Quantile Regression with Censored Data ». In : *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 54.2, p. 252–259. url : http://www.jstor.org/stable/25050878.

Dorfman, A. H. et P. Hall (1993). « Estimators of the finite population distribution function using nonparametric regression ». In : *Annals of Statistics* 21, p. 1452–1475.

Efron, B. (1967). « The Two Sample Problem with Censored Data ». In : *Proc. 5th Berkeley Symp.* 4, p. 831–853.

— (1991). « Regression percentiles using asymmetric squared error loss ». In : *Statistica Sinica* 1.1, p. 93–125. url : http://www.jstor.org/stable/24303995.

Falk, M. (1983). « Relative efficiency and deficiency of kernel type estimators of smooth distribution functions ». In : *Statistica Neerlandica* 37.2, p. 73–83. url : http://dx.doi.org/10.1111/j.1467-9574.1983.tb00802.x.

— (1984). « Relative Deficiency of Kernel Type Estimators of Quantiles ». In : *The Annals of Statistics* 12.1, p. 261–268. url : http://www.jstor.org/stable/2241047.

Fitzgerald, J. (1991). « Welfare durations and the marriage marker: evidence from the Survey of Income and Program Participation ». In : *Journal Human Resource* 3.26, p. 545–61.

Gonzalez-Manteiga, W. et C. Cadarso-Suarez (1994). « Asymptotic properties of a generalized kaplan-meier estimator with some applications ». In : *Journal of Nonparametric Statistics* 4.1, p. 65–78. url : http://dx.doi.org/10.1080/10485259408832601.

Hall, P., R. C. L. Wolff et Q. Yao (1999). « Methods for Estimating a Conditional Distribution Function ». In : *Journal of the American Statistical Association* 94.445, p. 154–163. url : http://www.jstor.org/stable/2669691.

Hu, Y. et G. Ridder (2012). « Estimation of nonlinear models with mismeasured regressors using marginal information ». In : *Journal of Applied Econometrics* 27.3, p. 347–385. url : http://dx.doi.org/10.1002/jae.1202.

Johnson, A. A., F. J. Breidt et J. D. Opsomer (2008). « Estimating Distribution Functions from Survey Data using Nonparametric Regression ». In : *Journal of Statistical Theory and Practice* 2, p. 419–431.

Kaplan, E. L. et P. Meier (1958). « Nonparametric Estimation from Incomplete Observations ». In : *Journal of the American Statistical Association* 53.282, p. 457–481. url : http://www.jstor.org/stable/2281868.

Kardaun, O. (1983). « Statistical survival analysis of male larynx-cancer patients - a case study ». In : *Statistica Neerlandica* 37.3, p. 103–125.

Kokic, P., R. Chambers, J. Breckling et S. Beare (1997). « A Measure of Production Performance ». In : *Journal of Business and Economic Statistics* 15.4, p. 445–451. url : http://www.jstor.org/stable/1392490.

Kuk, A. Y. C. (1988). « Estimation of Distribution Functions and Medians Under Sampling with Unequal Probabilities ». In : *Biometrika* 75.1, p. 97–103.

Li, G., R. C. Tiwari et M. T. Wells (1996). « Quantile Comparison Functions in Two-Sample Problems, With Application to Comparisons of Diagnostic Markers ».

In : *Journal of the American Statistical Association* 91.434, p. 689–698. URL : http://www.jstor.org/stable/2291664.

Li, K.-C. (1991). « Sliced Inverse Regression for Dimension Reduction ». In : *Journal of the American Statistical Association* 86.414, p. 316–327. URL : http://www.jstor.org/stable/2290563.

Lombardia, M.-J., Gonzalez-Manteiga, W. et W. Prada-Sanchez (2004). « Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function ». In : *Journal of Nonparametric Statistics* 16.1-2, p. 63–90.

Mukarjee, H. et S. Stern (1994). « Feasible Nonparametric Estimation of Multiargument Monotone Functions ». In : *Journal of the American Statistical Association* 89.425, p. 77–80. URL : http://www.jstor.org/stable/2291202.

Mukhopadhyay, P. (2001). *Topics in survey sampling*. Lecture notes in statistics. Springer. ISBN : 9780387951089.

Newey, W. K. et J. L. Powell (1987). « Asymmetric Least Squares Estimation and Testing ». In : *Econometrica* 55.4, p. 819–847. URL : http://www.jstor.org/stable/1911031.

Padget, W. J. et L. A Thombs (1988). « A smooth nonparametric quantile estimator from right-censored data ». In : *Statistics and Probability Letters* 7.2, p. 113–121. URL : http://www.sciencedirect.com/science/article/pii/0167715288900351.

Padgett, W. J. (1986). « A Kernel-Type Estimator of a Quantile Function From Right-Censored Data ». In : *Journal of the American Statistical Association* 81.393, p. 215–222. URL : http://www.jstor.org/stable/2287993.

Pfefferman, D. et C. R. Rao (2009). *Sample surveys: design, methods and applications*. Handbook of statistics. Elsevier. ISBN : 9780444531247.

Poiraud-Casanova, S. et C. Thomas-Agnan (1998). « Quantiles conditionnels ». In : *Journal de la société française de statistique* 139.4, p. 31–44. URL : http://eudml.org/doc/199555.

Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press. URL : http://www.sciencedirect.com/science/article/pii/B9780125640206500025.

Rao, J. N. K. (2003). *Small area estimation*. Wiley series in survey methodology. New York : John Wiley et Sons.

Rao, J. N. K., J. G. Kovar et H. J. Mantel (1990). « On estimating distribution functions and quantiles from survey data using auxiliary information ». In : *Biometrika* 77.2, p. 365–375. DOI : 10.1093/biomet/77.2.365.

Robertson, T., F. T. Wright et R. Dykstra (1988). *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley. ISBN : 9780471917878. URL : https://books.google.fr/books?id=sqZfQgAACAAJ.

Rueda, M. M., A. Arcos, M. D. Martínez-Miranda et Y. Román (2004). « Some improved estimators of finite population quantile using auxiliary information in sample surveys ». In : *Computational Statistics and Data Analysis* 45.4, p. 825–848.

Salvati, N., H. Chandra et R. Chambers (2010). « Model-based direct estimation of small area distributions ». In : *Centre for Statistical and Survey Methodology, working paper*.

Salvati, N., H. Chandra, M. G. Ranalli et R. Chambers (2010). « Small area estimation using a nonparametric model-based direct estimator ». In : *Journal of Computational Statistics and Data Analysis* 54, p. 2159–2171.

Salvati, N., M. G. Ranalli et M. Pratesi (2010). « Small area estimation of the mean using nonparametric M-quantile regression: a comparison when a linear mixed model does not hold ». In : *Journal of Statistical Computation and Simulation*, p. 1–20.

Samanta, M. (1989). « Non-parametric estimation of conditional quantiles ». In : *Statistics and Probability Letters* 7.5, p. 407–412. URL : http://EconPapers.repec.org/RePEc:eee:stapro:v:7:y:1989:i:5:p:407-412.

Van Keilegom, I. et N. Veraverbeke (1996). « Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles ». In : *Communications in Statistics - Theory and Methods* 25, p. 2251–2265. URL : http://dx.doi.org/10.1080/03610929608831836.

Wang, S. et A. H. Dorfman (1996). « A new estimator of the finite population distribution function ». In : *Biometrika* 83, p. 639–652.

Welsh, A. H. (1996). « Robust Estimation of Smooth Regression and Spread Functions and their Derivatives ». In : *Statistica Sinica* 6.2, p. 347–366. URL : http://www.jstor.org/stable/24306020.

Yao, Q. et H. Tong (1996). « Asymmetric least squares regression estimation: A nonparametric approach ». In : *Journal of Nonparametric Statistics* 6.2-3, p. 273–292. URL : http://dx.doi.org/10.1080/10485259608832675.

3 — Données censurées multivariées

Nous nous plaçons dans ce chapitre dans le cadre des données censurées multivariées, qui correspond au cas où l'on considère plusieurs événements d'intérêt. Ce cadre était celui de mon doctorat. Une revue des problèmes rencontrés et modèles utilisés dans ce contexte peut être trouvée dans Hougaard 2000.

L'appellation *données censurées multivariées* regroupe différents cas. On peut avoir avoir affaire à des événements répétés pour un même sujet, comme c'est le cas lorsqu'un observe plusieurs épisodes d'une maladie chronique pour un patient. Ce peut être aussi l'apparition d'événements de types différents pour un individu, ces types étant non exclusifs au sens où ces événements peuvent tous être observés si la durée de surveillance est suffisante. En cancérologie, on peut citer comme événements non exclusifs la récidive du cancer sur le même site, une récidive sur un site différent et l'apparition de métastases. Ce terme regroupe aussi le cas où l'on s'intéresse aux délais d'apparition d'un même type d'événement pour des sujets apparentés (fratrie, couple, cage...). Par exemple, Clayton et Cuzick 1985 considèrent les âges de décès par accident cardiaque du père et du fils. Dans ce dernier cas, c'est plutôt l'étude de l'association entre les deux événements qui présente de l'intérêt. Enfin, un dernier cas qui a connu un fort développement ces dernières années est celui des risques concurrents : un et seul événement parmi plusieurs événements concurrents potentiels est observé pour chaque individu : celui qui survient le premier.

Au cours de mon DEA et de mon doctorat, je me suis intéressée à la comparaison de deux groupes dans le cas d'événements répétés ou multiples non exclusifs et ai proposé une nouvelle famille de tests de rangs non paramétriques pour comparer deux distributions multivariées en présence de censure à droite [1]. Les modèles de régression associés à deux de ces tests ont ensuite été décrits puis généralisés à la prise en compte de plusieurs facteurs de risque. Ils permettent l'estimation de risques relatifs via une vraisemblance partielle analogue à celle de Cox 1972.

En collaboration avec Gérard Derzko, biostatisticien chez Sanofi-Synthélabo, nous avons tenté dans [4] d'apporter un peu de clarté dans la terminologie plutôt confuse qui est utilisée dans les articles traitant des risques concurrents. En particulier, nous avons montré que les données observées peuvent relever de deux modèles de sélection de base et avons développé pour ces deux modèles des algorithmes simples permettant l'estimation

non paramétrique des (sous-)distributions décrivant l'apparition des événements au cours du temps. Ces modèles peuvent être combinés en modèles complexes et les algorithmes proposés se généralisent sans difficulté à ces cas. Ce sera l'objet de la section 3.1.

D'autre part, dans le domaine médical, il est fréquent d'observer des événements répétés, qui peuvent être interrompus aléatoirement par un événement terminal ou par une censure. Avec Gérard Derzko, nous avons proposé dans [5] une procédure d'estimation non paramétrique de l'incidence cumulée de ces événements, à un rang donné puis totale, tout d'abord uniquement en présence de censure à droite, puis en présence de censure à droite et de troncature, c'est-à-dire en présence d'un événement terminal qui vient interrompre la séquence des événements répétés. La méthode, présentée dans la section 3.2, utilise les algorithmes développés dans [4]. Elle fournit également des estimateurs non paramétriques des prévalences associées à des événements de rang donné.

Dans le cadre de l'encadrement de la thèse de Serge Somda, en codirection avec Thomas Filleron, méthodologiste-biostatisticien à l'Institut Claudius Regaud, nous nous sommes intéressés à la méthodologie de la surveillance post-thérapeutique en oncologie : détermination de la durée optimale de surveillance en tenant compte des types concurrents de récidives auquel le patient est à risque et programmation optimale des visites de contrôle en utilisant un modèle de Markov homogène multi-états. Ce sera l'objet de la section 3.3, qui présente les articles [9], [11], [12], [13] et la prépublication [23].

3.1 Risques concurrents avec censure à droite

Cette section reprend les principaux développements de l'article [4].

3.1.1 Motivation

Le problème des risques concurrents est apparu à l'origine dans le domaine de la santé publique et des statistiques de mortalité : il s'agissait principalement d'estimer les courbes d'incidence de décès correspondant à chaque cause de décès et de prédire comment elles se modifieraient dans le cas où l'une des causes serait éradiquée. Le terme de risques concurrents désignait au départ l'étude d'événements fatals de différents types, mais ce terme a été étendu à tout problème où l'on peut distinguer différents types de défaillances, fatals ou non. Par exemple, en cardiologie, une population à risque d'accidents cardio-vasculaires peut manifester plusieurs types d'accidents : infarctus du myocarde, accident vasculaire cérébral... En cancérologie, on peut être intéressé par le temps écoulé jusqu'à la survenue du premier événement parmi plusieurs événements non mortels (récidive du cancer, apparition de métastases...). Ces événements peuvent être étudiés indistinctement en les considérant comme des réalisations d'une seule et même variable délai, mais il est encore plus intéressant d'appréhender leur comportement spécifique. Bien entendu, la censure à droite, due aux exclus-vivants ou aux perdus de vue, vient encore compliquer le phénomène en masquant l'observation des événements d'intérêt.

Dans la situation de base où l'étude porte sur les délais d'apparition d'un événement d'un type donné en présence d'une censure à droite indépendante de l'événement, Kaplan et Meier 1958 (voir 1.1) fournit un estimateur non paramétrique du maximum de vraisemblance de la fonction de survie de cet événement. Cette méthode est la plus connue des praticiens et est incluse dans la plupart des logiciels du commerce ; aussi est-elle utilisée parfois sans discernement dans toutes sortes de conditions expérimentales complexes, incluant plusieurs types d'événements et de censure. Les utilisateurs considèrent alors que tous les événements en compétition avec l'événement d'intérêt se

comportent comme une censure vis-à-vis de cet événement, en plus de la vraie censure : cette approche n'est valide qu'en supposant l'indépendance des événements entre eux, condition qui paraît d'évidence trop forte dans de nombreuses situations. L'emploi abusif de la procédure de Kaplan-Meier a été discutée par plusieurs auteurs (voir par exemple Gooley et al. 1999) et plusieurs méthodes ont été comparées (voir par exemple Arriagada et al. 1992).

Bien qu'il y ait beaucoup de confusion dans la terminologie, l'approche consistant à utiliser la méthode de Kaplan-Meier en cas de risques concurrents en considérant les autres événements comme de la censure est connue sous le nom de méthode des variables latentes. Une autre méthode d'estimation pour traiter des situations où la censure est indépendante des différents types d'événements d'intérêt, qui ne sont pas supposés indépendants entre eux, a été proposée par Kalbfleisch et Prentice 1980. Cette méthode est peu connue, peu utilisée en pratique, et absente de la plupart des logiciels [1].

Le contexte général de l'article [4] est celui où n unités statistiques sont présentes et où l'unité i ($1 \leq i \leq n$) fournit un résultat constitué d'un couple (Y_i,Δ_i) ; Δ_i est la réalisation d'une v.a. discrète Δ qui catégorise les événements et Y_i est la réalisation d'une v.a. continue positive Y (en pratique le temps écoulé depuis une origine fixe jusqu'à la survenue d'un événement quelle que soit sa nature Δ).

L'article [4] a pour objet de montrer que, dans le contexte général fixé, les données observées peuvent relever de deux modèles différents de sélection de base. L'estimation de Kaplan-Meier est pertinente pour un modèle de sélection par valeur minimale avec indépendance des délais. L'estimateur de Kalbfleisch et Prentice 1980 est toujours valide et correspond à un modèle de mélange censuré. Curieusement, l'idée d'un tel modèle apparaît rarement et seulement récemment dans la littérature (voir Betensky et Schoenfeld 2001) et la relation de ce modèle avec l'estimateur de Kalbfleisch et Prentice paraît ignorée. Nous développons pour ces deux modèles de base des algorithmes simples permettant l'estimation non paramétrique des (sous-)distributions décrivant l'apparition des événements au cours du temps. Ces modèles peuvent être combinés en modèles complexes et les algorithmes proposés se généralisent sans difficulté à ces cas.

Naturellement, les estimations obtenues par différentes méthodes sont différentes et induisent des conclusions et des interprétations différentes des données expérimentales. Le problème essentiel consiste alors à choisir le modèle sous-jacent de production des données. Malheureusement, ces modèles ne peuvent être distingués au moyen de tests d'hypothèses, comme cela a été remarqué très tôt par Tsiatis 1975. Cependant, bien qu'il n'y ait pas de méthode universelle pour ce choix, les conditions de l'expérimentation et la nature des événements étudiés permettent souvent de l'effectuer *a priori*.

La section 3.1.2 présente les deux modèles de sélection de base dans le cas le plus simple de deux événements en compétition sans censure. Dans la section 3.1.3, nous généralisons ces modèles au cas d'un nombre quelconque d'événements en compétition avec censure à droite. Nous montrons que ces modèles conduisent à deux estimateurs différents de l'incidence cumulée : l'estimateur de Kaplan-Meier et l'estimateur de Kalbfleisch et Prentice, et les algorithmes permettant d'obtenir ces estimateurs sont présentés. Si ces algorithmes n'ont pas grand intérêt dans les cas simples (car les estimateurs correspondants peuvent être obtenus autrement), leur utilité apparaît dans le cas de dispositifs complexes, tels ceux présentés à fin de la section 3.1.3 page 57. Nous expli-

1. En tout cas à la date de publication de l'article [4]. A noter que depuis 2011, cette estimation est disponible dans le logiciel `R` grâce à la fonction `cuminc` du paquet `cmprsk` (voir Gray 2011).

quons ensuite en section 3.1.4 comment générer des échantillons correspondant aux deux mécanismes de base dans les études de simulation et deux échantillons simulés selon les deux mécanismes sont analysés par les deux méthodes, pour illustrer l'erreur commise dans l'estimation en supposant un modèle inapproprié. Enfin, la section 3.1.5 s'intéresse au choix pratique du mécanisme de sélection.

3.1.2 Deux événements en compétition sans censure

Les mécanismes de sélection

Notations

Nous avons modifié les notations de l'article [4] pour les rendre cohérentes avec le reste du document. Considérons deux types d'événements en compétition. Les données observées constituent un n-échantillon marqué : $(Y_i,\Delta_i)_{i=1}^n$, où Y désigne le délai d'apparition d'un événement, variable continue de fonction de répartition $U(t) = P(Y < t)$ et où Δ, le type de l'événement, ou encore la marque, est une variable qualitative prenant ses valeurs dans $\{0,1\}$.

Notons $U_0(t) = P(Y < t,\Delta = 0)$ et $U_1(t) = P(Y < t,\Delta = 1)$ les incidences cumulées observées pour chaque type d'événement. Ces fonctions sont des sous-distributions. On a de façon évidente $U(t) = U_0(t) + U_1(t)$. Notons les distributions conditionnelles correspondantes $F_0(t) = P(Y < t \mid \Delta = 0)$ et $F_1(t) = P(Y < t \mid \Delta = 1)$.

Ce n-échantillon marqué peut être vu comme le résultat d'une sélection pouvant provenir de deux mécanismes différents, que nous allons détailler ci-après.

Le mécanisme de sélection par minimum (MSM)

La variable Y observée peut être considérée comme le minimum de deux variables T_0 et T_1, dénommées parfois dans la littérature « variables latentes » car non complètement observables. Cela correspond aux notations d'Efron (utilisées classiquement pour modéliser le phénomène de censure à droite d'un événement) :

$$\begin{aligned} Y &= \min(T_0,T_1), \\ \Delta &= j \text{ si } Y = T_j. \end{aligned}$$

En ajoutant l'hypothèse d'indépendance des variables T_0 et T_1 (hypothèse non testable), on obtient la relation suivante :

$$\begin{aligned} P(Y \geq t) = P(\min(T_0,T_1) \geq t) = P(T_0 \geq t)P(T_1 \geq t), \\ \text{qui peut s'écrire } 1 - U(t) = (1 - F_0(t))(1 - F_1(t)), \end{aligned} \tag{3.1}$$

qui est la condition caractérisant ce mécanisme de sélection.

Une image simple pour décrire ce mécanisme est celle de deux athlètes que l'on fait courir et l'on retient le temps du gagnant, ainsi que son identité. Ce mécanisme est le seul possible si l'une des marques désigne la censure aléatoire à droite, classiquement supposée indépendante de l'événement d'intérêt. Il correspond dans le cas de deux événements en compétition à des événements pouvant se réaliser pour chaque individu de façon certaine, mais qui ne sont éventuellement pas observés, par manque de temps. Des cas pratiques de deux événements certains et indépendants pouvant affecter des individus sont difficiles à trouver, ce qui explique que ce modèle a été souvent critiqué car correspondant à une situation irréaliste.

Si l'on ne fait pas l'hypothèse d'indépendance de T_0 et T_1, supposer que la variable observée Y s'obtient par minimum n'est plus exploitable pour une estimation non paramétrique et on peut alors considérer que l'échantillon relève d'un mécanisme de mélange.

Le mécanisme de mélange (MM)

Une autre approche possible est de considérer que la variable Y provient d'un mélange de distribution : deux types de délais coexistent dans l'échantillon en proportion $p_o = P(\Delta = 0)$ et $p_1 = 1-p_0$, fixées au départ et donc indépendantes du temps. Ce mécanisme ne peut donc pas être supposé si une des marques désigne la censure car la censure est un processus dynamique. L'approche en termes de mélange a été proposée par d'autres auteurs (voir par exemple BETENSKY et SCHOENFELD 2001).

La relation $U(t) = U_0(t) + U_1(t)$ s'écrit pour ce mécanisme, les proportions p_0 et p_1 des deux types étant fixées :

$$\begin{aligned} U(t) &= F_0(t)P(\Delta = 0) + F_1(t)P(\Delta = 1) \\ &= F_0(t)p_0 + F_1(t)(1 - p_0). \end{aligned}$$

Ce n'est pas le cas dans le MSM, où p_0 et p_1, si on voulait les écrire, seraient des fonctions de t. En effet, dans le MM, la différenciation de la formule ci-dessus donne :

$$dU(t) = P(\Delta = 0)dF_0(t) + P(\Delta = 1)dF_1(t),$$

alors que pour le MSM, la différenciation de la condition caractéristique (3.1) conduit à :

$$dU(t) = (1 - F_1(t))dF_0(t) + (1 - F_0(t))dF_1(t),$$

qui met clairement en évidence le caractère soit constant, soit fonction du temps de la décomposition linéaire de $dU(t)$ en fonction de $dF_0(t)$ et $dF_1(t)$ dans les deux modèles. De ce fait, une même paire (F_0,F_1) ne peut jamais être solution à la fois de l'une et l'autre équation.

Dans le MM, on peut toujours, comme dans le cas du MSM, associer deux variables aléatoires T_0 et T_1 aux fonctions F_0 et F_1 respectivement. Mais contrairement au cas du MSM, ces deux variables ne sont plus indépendantes, en tant que composantes d'un mélange, et elles n'ont plus d'interprétation en termes de variables latentes. La notation suivante pour la variable Y issue du mélange peut alors être utile : $Y = T_0 \cup_{p_0} T_1$, ce qui signifie, en posant $\tau(t) = [t,t+dt[$, que $P(T_0 \cup_{p_0} T_1 \in \tau(t)) = P(T_0 \in \tau(t)) + P(T_1 \in \tau(t))$.

L'image à retenir pour ce mécanisme est le suivant : on tire d'abord au sort l'athlète (selon une loi de Bernoulli de probabilité p_0) puis on le fait courir et on note son temps et son identité. Ce mécanisme correspond à des causes de mortalité exclusives. Il est parfois dénommé dans la littérature mécanisme des « causes spécifiques ». Il est dans la plupart des cas plus réaliste que le mécanisme de la sélection par minimum qui suppose l'indépendance des événements. Un exemple typique de ce mécanisme est le cas de l'étude de la mortalité humaine, les différents types correspondant à des causes exclusives de décès.

Estimation non paramétrique dans le cas du mécanisme de sélection par minimum

Les fonctions que l'on souhaite estimer dans ce cas sont les incidences cumulées correspondant à chaque événement, c'est-à-dire les fonctions de répartition conditionnelles F_0 et F_1. En effet, les événements étant supposés indépendants, cela a du sens de s'intéresser à la fonction de répartition d'un événement d'un type donné. Ces fonctions ne sont pas complètement observables. Par contre, les sous-distributions $U_0(t)$, $U_1(t)$ et leur somme $U(t)$ sont observables. Ces dernières peuvent donc être estimées de façon empirique sur l'échantillon.

T_0 et T_1 étant supposées indépendantes, on a la condition (3.1) :

$$1 - U(t) = (1 - F_0(t))(1 - F_1(t)).$$

En imposant de plus la condition de non-informativité suivante :

$$\left(\frac{\partial U_0}{\partial F_1}\right)(t) = \left(\frac{\partial U_1}{\partial F_0}\right)(t) \equiv 0, \tag{3.2}$$

c'est-à-dire qu'une variation de U_0 implique une variation de F_0 seulement et qu'une variation de U_1 implique une variation de F_1 seulement, on obtient (voir démonstration en annexe de [4]) le système d'équations différentielles suivant en $F_0(t)$ et $F_1(t)$:

$$\begin{cases} dF_0(t) &= \dfrac{dU_0(t)}{1-F_1(t)} \\ dF_1(t) &= \dfrac{dU_1(t)}{1-F_0(t)} \end{cases} \tag{3.3}$$

soit

$$\begin{cases} dF_0(t) = & (1-F_0(t))\,\dfrac{dU_0(t)}{1-U(t)} \\ dF_1(t) = & (1-F_1(t))\,\dfrac{dU_1(t)}{1-U(t)} \end{cases} \tag{3.4}$$

Remarquons que la condition de non-informativité 3.2 est analogue à la condition qui justifie l'usage d'une vraisemblance partielle — excluant les contributions de la censure — dans le cadre traditionnel de l'estimation par maximisation de vraisemblance. Cette condition n'est autre qu'une condition d'exhaustivité de U_0 pour F_0 et de U_1 pour F_1 au sens de Halmos et Savage 1949.

Le système obtenu peut s'écrire sous la forme :

$$\begin{cases} \dfrac{dF_0(t)}{1-F_0(t)} = & \dfrac{dU_0(t)}{1-U(t)} \\ \dfrac{dF_1(t)}{1-F_1(t)} = & \dfrac{dU_1(t)}{1-U(t)} \end{cases} \tag{3.5}$$

qui correspond à l'égalité de deux risques, obtenue du fait de l'indépendance des deux délais d'événement : les membres de gauche sont parfois appelés dans la littérature « risques nets », alors que les membres de droite sont appelés « risques bruts » ou « risques apparents », et l'on voit qu'une résolution formelle est possible et que de plus la partie droite des équations est observable, donc estimable empiriquement. La solution de ce système est :

$$\begin{cases} F_0(t) &= 1-\exp\left(-\displaystyle\int_0^t \frac{dU_0(x)}{1-U(x)}\right) \\ F_1(t) &= 1-\exp\left(-\displaystyle\int_0^t \frac{dU_1(x)}{1-U(x)}\right) \end{cases}$$

Le remplacement de U, U_0 et U_1 dans ces solutions par leurs estimateurs empiriques conduit aux estimateurs de Harrington et Fleming (voir page 16) des fonctions F_0 et F_1.

Nous choisissons plutôt d'obtenir des estimations de F_0 et de F_1 à l'aide d'un algorithme. L'échantillon total est ordonné par valeurs croissantes des temps observés. On note cet échantillon ordonné $(y_i,\delta_i)_{i=1}^n$. L'ensemble $\{y_1,y_2,\ldots,y_n\}\times\{0,1\}$ est probabilisé en donnant la probabilité $\frac{1}{n}$ à chacun des n points observés. On cherche des estimateurs $\widehat{F}_0$ et $\widehat{F}_1$ de F_0 et F_1 qui vérifient le système (3.4) à chaque temps t. Bien entendu, ces

estimateurs dépendent de n, taille de l'échantillon, mais l'indice n sera omis pour ne pas alourdir les notations.

Cela conduit à :

$$\widehat{F}_j(y_i) - \widehat{F}_j(y_{i-1}) = \left(1 - \widehat{F}_j(y_{i-1})\right) \widehat{\lambda}_{ji}, \quad j = 0,1, \tag{3.6}$$

$$\text{avec } \widehat{\lambda}_{ji} \quad = \quad \frac{\widehat{U}_j(y_i) - \widehat{U}_j(y_{i-1})}{1 - \widehat{U}(y_{i-1})},$$

avec comme conditions initiales $\widehat{F}_j(0) = 0$, $j = 0,1$.

En prenant comme estimateurs de U et des U_j, $j = 0,1$, les estimateurs empiriques correspondants, on obtient

$$\widehat{\lambda}_{ji} = \frac{\frac{1}{n}\mathbb{1}(\delta_i = j)}{1 - \sum_{1}^{i-1} \frac{1}{n}} = \frac{\mathbb{1}(\delta_i = j)}{n - i + 1}.$$

En posant $\widehat{S}_j(y_i) = 1 - \widehat{F}_j(y_i)$, $j = 0,1$, on obtient immédiatement à partir de l'équation (3.6) :

$$\widehat{S}_j(y_i) = \widehat{S}_j(y_{i-1})(1 - \widehat{\lambda}_{ji}),$$

avec comme conditions initiales $\widehat{S}_j(0) = 1$, $j = 0,1$. Par récurrence sur i et en étendant le résultat à une fonction en escalier sur $\mathbb{R}$, on obtient alors

$$\widehat{S}_j^{\text{KM}}(t) \quad = \quad \prod_{i:y_i \leq t} \left(1 - \frac{\mathbb{1}(\delta_i = j)}{n - i + 1}\right), \quad j = 0,1, \tag{3.7}$$

qui sont les estimateurs de Kaplan-Meier des fonctions de survie $1 - F_0(t)$ et $1 - F_1(t)$, obtenus en considérant que les deux événements se censurent mutuellement. Ces estimateurs ont la propriété de vérifier l'équation (3.1), qui correspond à la condition d'indépendance, à chaque temps, alors que les estimateurs de Harrington-Fleming ne la vérifie qu'asymptotiquement. Ces mêmes estimations peuvent être obtenues par l'algorithme 1, obtenu à partir du système (3.3) et dont l'intérêt apparaîtra plus loin.

Algorithme 1 MSM : cas de deux événements en compétition

L'échantillon total est ordonné par valeurs croissantes des temps : on le note $(y_i,\delta_i)_{i=1}^n$.
Initialisation : $\widehat{F}_j(0) = 0, \quad j = 0,1$.
Pour $i = 1$ à n, $j = 0,1$,

$$\widehat{F}_j(y_i) = \widehat{F}_j(y_{i-1}) + \frac{1}{n} \frac{\mathbb{1}(\delta_i = j)}{1 - \widehat{F}_{(1-j)}(y_{i-1})}.$$

La convergence de cet algorithme est assurée car il conduit aux estimations de Kaplan-Meier. L'intérêt de cet algorithme est qu'il s'étend à des modèles beaucoup plus complexes qui seront vus plus loin. Les estimations des risques instantanés et des densités sont également possibles.

Estimation non paramétrique dans le cas du modèle de mélange

Dans le cas très simple du mélange de deux distributions non censurées, les proportions p_0 et p_1 peuvent être estimées par les proportions correspondantes observées dans l'échantillon. Dans le cas du mélange, les fonctions d'intérêt sont les incidences cumulées U_0 et U_1, qui correspondent aux incidences de chaque type d'événement, en présence du risque associé à l'autre type d'événement. On ne s'intéresse pas dans ce cas à F_0 et F_1, qui correspondraient à des incidences d'un événement d'un type donné, considéré comme le seul possible. Les fonctions U_0 et U_1 sont observables et peuvent être estimées de façon empirique par :

$$\widehat{U}_k^{Emp}(t) = \frac{1}{n} \sum_{i:y_i \leq t} \mathbb{1}(\delta_i = k), \quad k = 0,1. \tag{3.8}$$

3.1.3 Plusieurs événements en compétition avec censure

La situation va maintenant se compliquer de la façon suivante : on est en présence de plusieurs événements concurrents et d'un mécanisme de censure à droite. Un seul délai est observé pour chaque individu : le délai d'un des événements d'intérêt ou le délai de censure.

On est toujours en présence d'un échantillon marqué : $(Y_i,\Delta_i)_{i=1}^n$, avec Δ qui prend ses valeurs dans $\{0,1,\ldots,p\}$. La marque 0 désigne la censure.

Comme précédemment, cet échantillon peut être vu comme le résultat d'un des deux mécanismes de sélection suivants :

- Mécanisme de sélection par minimum (MSM) : c'est la généralisation du mécanisme vu dans le cas de deux événements. On considère que le temps Y observé est obtenu comme minimum de $p+1$ variables latentes que l'on supposera indépendantes, et dont l'une, T_0, est le délai de censure.
- Mécanisme de mélange censuré (MMC) : le temps observé est le minimum d'un délai de censure et d'une variable aléatoire dont la distribution est un mélange à p composantes.

Ces deux mécanismes de sélection conduisent à des estimateurs différents des fonctionnelles d'intérêt. Les deux algorithmes correspondants vont être présentés.

Mécanisme de sélection par minimum : cas général

Le couple (Y,Δ) est obtenu de la façon suivante :

$$\begin{aligned} Y &= \min(T_0,T_1,T_2,\ldots,T_p) \\ \Delta &= j \text{ si } Y = T_j \end{aligned}$$

et les variables T_j sont supposées indépendantes deux à deux. Cette hypothèse forte n'est malheureusement pas testable à cause d'un problème d'identifiabilité (voir Tsiatis 1975).

La condition caractérisant le mécanisme de sélection s'écrit :

$$1 - U(t) = \prod_{j=0}^{p} (1 - F_j(t)) \,.$$

Le système d'équations (3.3) obtenu dans le cas de deux événements sans censure, qui découle de l'hypothèse de non-informativité, se généralise aisément au cas de $p+1$ variables :

$$dF_j(t) = \frac{dU_j(t)}{\prod_{k \neq j}(1 - F_k(t))} = (1 - F_j(t))\frac{dU_j(t)}{1 - U(t)}, \quad j = 1,\ldots,p.$$

En cherchant des estimateurs qui vérifient ce système à chaque temps, on aboutit comme précédemment aux estimateurs de Kaplan-Meier pour les $F_j(t)$, $j = 0,1,\ldots,p$, en considérant les autres variables comme des censures, en plus de la vraie censure T_0.

L'algorithme 1 se généralise très facilement et on obtient l'algorithme 2.

Algorithme 2 MSM : cas général

L'échantillon total est ordonné par valeurs croissantes des temps : on le note $(y_i,\delta_i)_{i=1}^n$.
Initialisation : $\widehat{F}_j(0) = 0$, $j = 0,\ldots,p$.
Pour $i = 1$ à n, $j = 0,\ldots,p$,

$$\widehat{F}_j(y_i) = \widehat{F}_j(y_{i-1}) + \frac{\mathbb{1}(\delta_i = j)}{n \prod_{k\neq j}(1 - \widehat{F}_k(y_{i-1}))}.$$

Bien entendu, les autres fonctionnelles d'intérêt (risques instantanés, densités) peuvent également être obtenues.

Mécanisme de mélange censuré

La symétrie des $p+1$ variables latentes du mécanisme précédent ne s'applique plus ici. Nous notons donc T_0 le délai de censure et $T_{1,k}$, $k = 1,\ldots,p$, les p délais des événements (théoriques et sans interprétation propre).

Ces délais $T_{1,k}$ ne sont pas indépendants entre eux conditionnellement à $\Delta \neq 0$ (en tant que composantes d'un mélange) ; par contre, on fait l'hypothèse classique d'indépendance du délai de censure T_0 avec chaque $T_{1,k}$.

Le mécanisme de sélection est le suivant :

$$\begin{aligned} Y &= \min(T_0, \bigcup_{\omega} T_{1,k}) \\ \Delta &= 0 \text{ si } Y = T_0 \\ \Delta &= k \text{ si } Y = T_{1,k}, \end{aligned} \tag{3.9}$$

où $T_1 = \bigcup_\omega T_{1,k}$ est le résultat d'un mélange, w désignant une loi de probabilité définissant la composition de ce mélange. L'image est la suivante : on tire au sort un athlète, on le fait courir, mais il est alors en compétition avec un autre athlète (la censure) qui peut arriver avant lui : on note le temps du gagnant et son identité.

Notons $\omega_k = P(\Delta = k \mid \Delta \neq 0)$, $F_1(t)$ la fonction de répartition de la variable T_1 et $F_{1,k}(t)$, $k = 1,\ldots,p$, les distributions conditionnelles correspondant aux composantes du mélange. On a :

$$F_1(t) = \sum_{k=1}^{p} F_{1,k}(t)\ \omega_k = \sum_{k=1}^{p} I_{1,k}(t). \tag{3.10}$$

Il n'est pas nécessaire d'estimer explicitement les w_k car les quantités qui nous intéressent ici sont les $I_{1,k}(t)$, incidences cumulées des événements de type k en présence des autres types, et non pas les composantes du mélange $F_{1,k}(t)$, qui correspondraient aux incidences des événements de type k, le type k étant le seul possible. Contrairement au cas de deux événements vu précédemment, ces sous-distributions $I_{1,k}(t)$ ne sont plus entièrement observables à cause de la censure.

La condition caractérisant le mécanisme de sélection s'écrit :

$$\begin{aligned} 1-U(t) &= 1-\sum_{k=1}^{p} U_{1,k}(t) - U_0(t) \\ &= \left(1-\sum_{k=1}^{p} I_{1,k}(t)\right)\,(1-F_0(t))\,, \end{aligned} \tag{3.11}$$

où $U_0(t) = P(Y < t,\Delta = 0)$ et $U_{1,k}(t) = P(Y < t,\Delta = k)$, $k = 1,\ldots,p$. Ces dernières quantités sont toujours observables, donc estimables empiriquement.

En ajoutant la condition d'exhaustivité des $U_{1,k}$ pour les $I_{1,k}$ et de U_0 pour F_0, on obtient, comme précédemment :

$$\left\{\begin{array}{rcl} dI_{1,k}(t) &=& \dfrac{dU_{1,k}(t)}{(1-F_0(t))}, \quad k=1,\ldots,p, \\ dF_0(t) &=& \dfrac{dU_0(t)}{\left(1-\sum_{k=1}^{p} I_{1,k}(t)\right)}. \end{array}\right. \tag{3.12}$$

Ce système peut se réécrire :

$$\left\{\begin{array}{rcl} dI_{1,k}(t) &=& (1-F_1(t))\,\dfrac{dU_{1,k}(t)}{(1-U(t))}, \quad k=1,\ldots,p, \\ dF_0(t) &=& (1-F_0(t))\,\dfrac{dU_0(t)}{(1-U(t))}, \end{array}\right. \tag{3.13}$$

et sa solution est :

$$I_{1,k}(t) = \int_0^t \frac{dU_{1,k}(x)}{1-U(x)}\,\exp\left(-\int_0^x \frac{dU_1(z)}{1-U(z)}\right).$$

Le remplacement de U, U_1 et $U_{1,k}$ dans la formule ci-dessus par leurs estimateurs empiriques produit pour $I_{1,k}$ l'estimateur d'Aalen-Johansen dont les propriétés mathématiques peuvent être trouvées dans le chapitre IV.4. de Andersen, Borgan et al. 1993.

Estimation

L'ensemble $\{y_1,y_2,\ldots,y_n\} \times \{0,1,2,\ldots,p\}$ est probabilisé par :

$$\begin{aligned} P(Y=t,\Delta=k) &= \frac{1}{n} \text{ si } \exists i : (t,k) = (y_i,\delta_i), \\ &= 0 \text{ sinon,} \end{aligned}$$

c'est-à-dire qu'à chaque point observé est affectée une probabilité $\frac{1}{n}$.

En utilisant (3.13), la discrétisation donne, pour $k = 1,\ldots,p$:

$$\widehat{I}_{1,k}(y_i) - \widehat{I}_{1,k}(y_{i-1}) = \left(1-\widehat{F}_1(y_{i-1})\right)\,\widehat{\lambda}_{1,ki} \tag{3.14}$$

$$\begin{aligned} \text{avec } \widehat{\lambda}_{1,ki} &= \frac{\widehat{U}_{1,k}(y_i) - \widehat{U}_{1,k}(y_{i-1})}{1-\widehat{U}(y_{i-1})} \\ &= \frac{\mathbb{1}(\delta_i = k)}{n-i+1}. \end{aligned}$$

L'estimation $\widehat{F}_1(t)$ de $F_1(t) = \sum_{k=1}^{p} I_{1,k}(t)$ s'obtient en sommant sur k les premières équations de (3.12) et on a aboutit alors au système suivant :

$$\begin{cases} dF_1(t) = & \dfrac{dU_1(t)}{1 - F_0(t)} \\ dF_0(t) = & \dfrac{dU_0(t)}{1 - F_1(t)} \end{cases}$$

qui est exactement le système (3.3). L'estimation qui en découle est donc l'estimation de Kaplan-Meier de la fonction de survie $S_1 = 1 - F_1$ (obtenue en pratique en regroupant tous les événements d'intérêt en un seul type). On aboutit alors, par récurrence sur i de la formule (3.14), à l'estimateur de l'incidence cumulée pour la cause k ($k = 1, \ldots, p$) proposé de façon heuristique par Kalbfleisch et Prentice 1980 :

$$\widehat{I}_{1,k}^{\mathrm{Pr}}(t) = \sum_{i:y_i \leq t} \frac{\mathbb{1}(\delta_i = k)}{n - i + 1} \quad \widehat{S}_1^{\mathrm{KM}}(y_i). \tag{3.15}$$

L'estimation de la fonction de survie $S_0 = 1 - F_0$ de la censure s'obtient comme celle de S_1 par Kaplan-Meier, en considérant tous les autres événements comme censurant la censure.

L'algorithme 3 permet d'obtenir les fonctions d'incidence cumulée et la fonction de répartition du délai de censure.

Algorithme 3 MMC

Initialisation : $\widehat{I}_{1,k}(0) = 0$, $k = 1, \ldots, p$, $\widehat{F}_0(0) = 0$.
Pour $i = 1$ à n, $k = 1, \ldots, p$,

$$\widehat{I}_{1,k}(y_i) = \widehat{I}_{1,k}(y_{i-1}) + \frac{\mathbb{1}(\delta_i = k)}{n\,(1 - \widehat{F}_0(y_{i-1}))},$$

$$\widehat{F}_0(y_i) = \widehat{F}_0(y_{i-1}) + \frac{\mathbb{1}(\delta_i = 0)}{n\left(1 - \sum_{k=1}^{p} \widehat{I}_{1,k}(y_{i-1})\right)}.$$

Généralisation des mécanismes MSM et MMC

Dans le mécanisme de mélange censuré, la censure peut également être le résultat d'un mélange (plusieurs causes exclusives de censure). Par exemple, en recherche clinique, on se trouve parfois amené à comparer, dans plusieurs groupes thérapeutiques où les survies sont similaires, les différentes causes de censure à droite (mauvaise tolérance, événement intercurrent, perte de vue...) : une répartition différente de ces causes de censure entre les groupes suggérerait un biais dans la comparaison des mortalités.

Un cas encore plus général est celui où la variable délai observée est le résultat d'une sélection par minimum de plusieurs variables latentes, chacune d'entre elles pouvant être le résultat d'un mélange. C'est le cas lorsque plusieurs types de censure (mélange de censure) sont en compétition avec plusieurs types d'événements (eux-mêmes mélange d'événements). Le cas du mécanisme de mélange censuré avec la censure résultat d'un

mélange envisagé au début de cette section correspond en fait au cas de deux variables latentes seulement. Supposons de façon générale qu'on ait affaire à J variables latentes, chacune résultat d'un mélange de n_j composantes. Le couple (Y,Δ) est alors obtenu ainsi :

$$\begin{aligned} Y &= \min(\bigcup_{\omega_1} T_{1,k}, \ldots, \bigcup_{\omega_J} T_{J,k}) \\ \Delta &= (j,k) \text{ si } Y = T_{j,k}. \end{aligned} \tag{3.16}$$

L'algorithme 4 permet d'estimer les fonctions d'incidence cumulées $I_{j,k}$, $j = 1,\ldots,J$, $k = 1,\ldots,n_j$.

Algorithme 4 Généralisation des MSM et MMC

L'échantillon total, ordonné par valeurs croissantes des temps, est noté $(y_i,\delta_i)_{i=1}^n$ (notons qu'ici δ_i désigne le couple $(j,k)_i$)).
Initialisation : $\widehat{I}_{j,k}(0) = 0,\ j = 1,\ldots,J,\ k = 1,\ldots,n_j$.
Pour $i = 1$ à n, $j = 1,\ldots,J$, $k = 1,\ldots,n_j$,

$$\widehat{I}_{j,k}(y_i) = \widehat{I}_{j,k}(t_{i-1}) + \frac{\mathbb{1}(\delta_i = (j,k))}{n \prod_{j' \neq j} \left(1 - \sum_{l=1}^{n_{j'}} \widehat{I}_{j',l}(y_{i-1})\right)}.$$

Les estimations des incidences cumulées des variables latentes $T_1,\ldots,T_J$ (variables résultant des J mélanges) s'obtiennent par la procédure de Kaplan-Meier, en groupant tous les événements du même mélange j et en considérant les autres événements comme des censures. L'estimateur obtenu vérifie :

$$\widehat{F}_j^{\mathrm{KM}}(t) = \sum_k \widehat{I}_{j,k}^{Pr}(t),$$

où les $\widehat{I}_{j,k}^{Pr}$ sont les estimateurs des incidences cumulées des n_j composantes du mélange j, obtenus par l'algorithme précédent, et généralisant l'estimateur (3.15).

D'autre part, on peut facilement vérifier que dans le cas général les estimateurs obtenus vérifient à chaque temps t la propriété vérifiée par les quantités théoriques, à savoir :

$$1 - \sum_{j,k} \widehat{U}_{j,k}^{Emp}(t) = \prod_j \left(1 - \sum_k \widehat{I}_{j,k}^{Pr}(t)\right),$$

où les $\widehat{U}_{j,k}^{Emp}$ sont les généralisations des estimateurs définis en (3.8).

3.1.4 Simulation d'un échantillon marqué

Principe

Sélection par minimum

Pour générer un n-échantillon marqué provenant d'une sélection par minimum, il faut générer $p+1$ variables aléatoires indépendantes (les variables latentes) selon des lois F_j, $j = 0,\ldots,p$: est retenu pour l'individu i le minimum des $p+1$ réalisations et le numéro du type correspondant.

Sélection par minimum et mélange

On se donne une loi de probabilité ω pour les types k, $k = 1, \ldots, p$, du mélange et des lois $F_{1,k}$ associées à chacun de ces types. On se donne également F_0, la loi de la censure.

Pour chaque individu i, on tire aléatoirement son type $\delta_i = k$ suivant la loi ω, puis on tire son temps $T_{1,k}$ selon la loi $F_{1,k}$ et son délai de censure T_0 suivant la loi F_0.

Est retenu le minimum de $T_{1,k}$ et T_0 avec l'indicateur de type correspondant (0 ou k).

Illustration

Nous avons généré un échantillon de taille 500 provenant d'un mécanisme de sélection par minimum : les deux durées d'événement suivent des lois exponentielles de paramètres respectifs 1 et 2 et elles sont censurées par un délai de censure de loi exponentielle de paramètre 1. Cet échantillon a été analysé par les deux méthodes : la méthode pertinente dans ce cas (algoMSM) et la méthode erronée (algoMMC).

De même, un échantillon de données de 500 individus provenant d'un mécanisme de mélange censuré a été généré : le mélange se compose de deux lois exponentielles de paramètres respectifs 1 et 2 en proportions égales, censurées par un délai de censure de loi exponentielle de paramètre 1. Cet échantillon a également été analysé par les deux méthodes.

La figure 3.1 montre les courbes d'incidences cumulées estimées dans les 4 cas. Pour visualiser l'erreur faite en postulant un mécanisme qui n'est pas le bon, nous avons tracé en traits pleins les courbes théoriques correspondant aux fonctions d'intérêt pour chaque mécanisme généré. Dans le cas des données provenant d'un MSM, il s'agit des fonctions de répartition marginales F_1 et F_2 des événements de chaque type. Pour les données provenant d'un MMC, il s'agit des incidences cumulées correspondant à chaque type d'événement $I_{1,1}$ et $I_{1,2}$, qui correspondent à des probabilités d'occurrence des événements de chaque type en présence des autres types et dont la somme donne la fonction de répartition de la variable issue du mélange.

On constate tout d'abord que dans le cas où on applique l'algorithme pertinent, les estimations s'ajustent tout à fait aux courbes théoriques. Par contre, l'utilisation de la « mauvaise méthode » conduit à des courbes très différentes, ce qui s'explique par le fait que l'on n'estime pas les mêmes fonctions. Sur cet exemple, l'analyse de l'échantillon MSM par l'algorithme MMC (en bas à gauche) ne conduit pas à des conclusions fondamentalement différentes (les deux courbes estimées sont simplement plus basses, mais les événements de type 2 ont, pour tous les temps, une incidence cumulée plus forte que celle des événements de type 1, comme les courbes théoriques). Par contre, l'analyse de l'échantillon MMC par l'algorithme MSM (en haut à droite) est plus trompeuse : en effet, les estimations de Kaplan-Meier des fonctions de répartition des deux événements se croisent, si bien qu'on serait amené à conclure que la survie pour l'événement de type 1 est meilleure que pour le type 2 au début, mais que cette tendance s'inverse après le point d'intersection, alors qu'en réalité l'incidence cumulée de l'événement de type 2 est toujours au dessus de celle de l'événement de type 1. Il est difficile de dire dans le cas général quelle erreur est la plus grave, mais il semble malgré tout plus prudent de ne pas considérer comme indépendants des événements qui ne le sont pas et donc d'utiliser l'algorithme MMC en cas de doute sur l'indépendance.

3.1.5 Choix du mécanisme de sélection

Comme nous l'avons montré sur les exemples précédents, le choix du mécanisme de sélection conduit à des estimations différentes. La question pratique principale reste donc de spécifier correctement le modèle de sélection, sachant que sa validité ne peut être

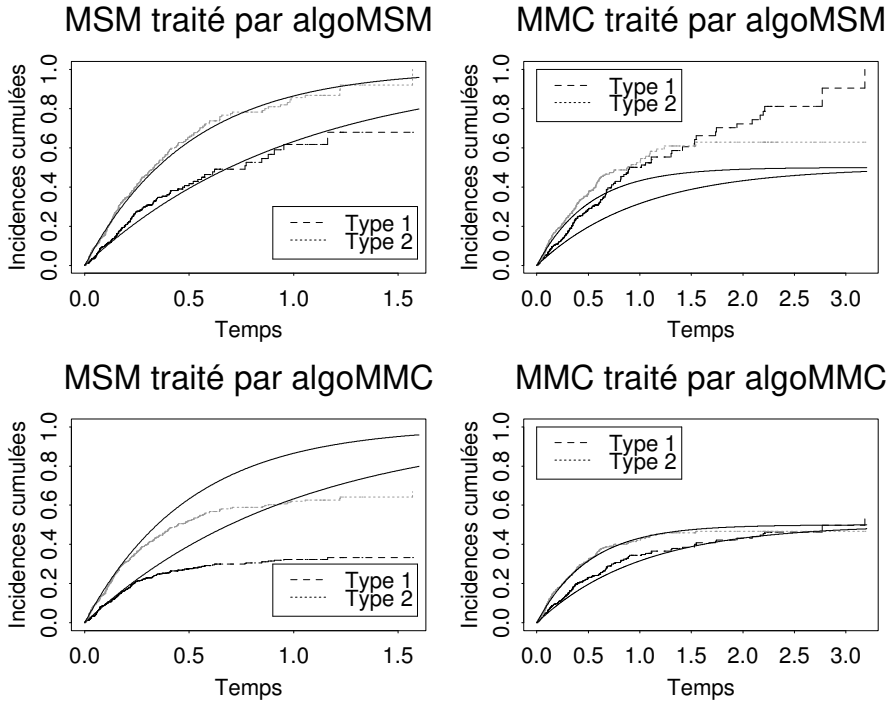


FIGURE 3.1 – Estimations des incidences cumulées pour deux échantillons MSM et MMC analysés par les deux algorithmes. Les courbes en traits pleins correspondent aux incidences cumulées théoriques.

testée. Il n'y a pas de réponse universelle à cette question, bien que les arguments pour ce choix existent souvent, spécialement en faveur du modèle de mélange censuré, qui ne suppose pas l'indépendance des événements. C'est typiquement le cas lorsque la liaison entre les variables d'intérêt est un phénomène « de nature », comme dans le schéma dit « à trois états », où un patient à risque peut soit devenir malade puis décéder, soit décéder directement, lorsqu'on ne s'intéresse qu'à estimer l'incidence du premier événement : un exemple de ce schéma est celui des accidents cardio-vasculaires (infarctus du myocarde, accident vasculaire cérébral, hémorragie), qui peuvent apparaître chez des patients à risque, et devenir éventuellement fatals, ou s'avérer fatals d'emblée. L'indépendance entre les occurrences d'un accident non fatal et du décès chez un patient est une supposition très peu raisonnable, car elle implique que la maladie et le décès puissent survenir chez un patient dans un ordre quelconque ! On peut difficilement concevoir une méthode d'estimation pour l'incidence de l'événement non fatal qui ne prendrait pas en compte l'existence de l'événement fatal. L'un des problèmes originels des risques concurrents (rappelé plus haut) équivaut sur cet exemple à vouloir prévoir l'incidence d'événements non fatals, dans l'hypothèse où une nouvelle technologie médicale viendrait supprimer le risque d'événement fatal. Cette incidence ne pourrait être estimée qu'en supposant un modèle à variables latentes avec indépendance des événements, mais ce modèle est ici a priori invalide ; au contraire, la considération d'un mélange d'événements cardio-vasculaires non fatals et fatals, qui ne suppose pas l'indépendance, semble refléter correctement le contexte.

La popularité du modèle des variables latentes vient très probablement de la facilité d'estimation des fonctions de survie correspondantes par la méthode de Kaplan-Meier,

présente dans tous les logiciels. D'autre part, ce modèle fournit le résultat le plus complet, à savoir la distribution multivariée des variables latentes, et l'utilisateur oublie facilement sous quelles conditions drastiques et irréalistes ce résultat est valide. Parallèlement, c'est l'absence de logiciel pour l'estimateur de Kalbfleisch et Prentice qui a certainement limité son utilisation pratique, alors que le mécanisme associé de mélange censuré est toujours valide, souvent plus réaliste, car prenant en compte les interrelations de nature entre les variables.

Remarquons enfin que la distinction entre les deux modèles de sélection n'a pas de conséquence en terme de tests non paramétriques d'hypothèse nulle, lorsqu'on souhaite comparer différents groupes indépendants. Le test du logrank (voir 1.3) ne dépend en effet que des risques empiriques apparents, qui sont identiques dans les deux modèles de sélection. Cette remarque ne fait qu'illustrer l'impossibilité de tester le choix du modèle de sélection.

3.2 Événements répétés avec censure et événement terminal

Cette section reprend les principaux éléments de l'article [5].

3.2.1 Motivation

Dans le domaine médical, il est fréquent d'observer des événements répétés chez un même patient On veut décrire comment ces événements apparaissent dans le temps, au moyen de ce que l'on appellera leur incidence, et aussi comparer ces courbes dans différents groupes. Le problème se complique du fait que l'observation statistique peut être interrompue de façon aléatoire : c'est le mécanisme de censure à droite. D'autre part, un événement terminal tel que le décès peut interrompre complètement le processus des répétitions, conduisant à une troncature des événements non mortels.

Dans cette situation, la méthode d'estimation non paramétrique usuelle consiste à ne considérer que la première occurrence du phénomène chez chaque patient et à utiliser la procédure d'estimation de Kaplan et Meier 1958 basée sur le mécanisme de censure à droite d'Efron. Cette procédure est bien adaptée au cas du décès, qui est un événement unique, mais elle implique dans le cas des événements répétés une perte d'information qui peut être considérable, à savoir tous les temps d'événements qui surviennent ultérieurement. En pratique, du fait de la non disponibilité de méthodes d'estimation non paramétriques appropriées, l'expérimentateur arrête l'observation après la première occurrence, alors même que l'objet de l'étude est l'estimation de la fréquence globale de survenue en fonction du temps, sans considération particulière pour le premier événement chez un patient donné.

Ces vingt dernières années[2], des modèles semi-paramétriques généralisant le modèle de Cox (voir 1.4) au cas des événements répétés ont été développés (voir Andersen et Gill 1982 ; Prentice et al. 1981 ; Wei et al. 1989). Comme dans le modèle de Cox usuel, l'accent est surtout mis dans ces modèles sur l'estimation des paramètres associés aux facteurs de risque plutôt que sur l'estimation de l'incidence cumulée des événements. D'autre part, les événements terminaux sont rarement pris en compte ; quand ils le sont (méthode de Wei et al. 1989), ils sont considérés comme des censures vis-à-vis des événements récurrents, approche qui nous paraît incorrecte du fait de la dépendance entre les événements récurrents et l'événement terminal. Dans un cadre non paramétrique, Lawless et Nadeau 1995 ont proposé un estimateur convergent de l'incidence globale (qu'ils appellent « fonction moyenne cumulative ») des événements récurrents dans le cas

2. Par rapport à la date de l'article, qui date de 2004.

où il n'y a pas d'événement terminal, mais seulement une censure aléatoire indépendante. Cook et Lawless 1997 ont étendu cet estimateur non paramétrique au cas où il y a des événements terminaux, qui peuvent être dépendants des événements récurrents, en présence de censure aléatoire indépendante des événements récurrents et terminaux ; les propriétés asymptotiques de cet estimateur ont été étudiées par Ghosh et Lin 2000. Cet estimateur est convergent mais il ne fournit d'estimation que pour l'incidence globale des événements répétés. En effet, hormis les modèles de Cox généralisés, aucune des méthodes de la littérature ne fournit à notre connaissance (au moins à la date de soumission de [5]) d'estimateur de l'incidence des événements qui se produisent à un rang donné dans la suite des événements récurrents. Or, l'incidence des événements rang par rang a un intérêt en soi, en particulier pour vérifier la cohérence des effets d'un traitement sur les différentes récurrences : en effet, dans certains cas, un phénomène de compensation peut exister, où un traitement retardera par exemple le premier événement d'un individu, mais aura l'effet inverse sur les événements ultérieurs. D'autre part, les estimations rang par rang des incidences cumulées permettent par différence le calcul de la prévalence associée à un événement de rang donné, qui présente un intérêt clinique certain. Menjoge 2003 a proposé un estimateur de l'incidence globale construit comme une somme d'estimateurs des incidences d'événements aux différents rangs ; malheureusement, ces estimateurs ne sont pas convergents.

Nous montrons dans la section 3.2.2 que ces situations peuvent être modélisées à l'aide d'un processus de comptage censuré. Nous développons dans la section 3.2.3 une procédure non paramétrique d'estimation applicable au cas d'un processus censuré, sans événement terminal ; puis dans la section 3.2.4, nous proposons une procédure qui s'étend au cas d'une troncature aléatoire à droite des événements récurrents par un événement terminal. Ces procédures font appel à des algorithmes très simples, qui découlent des mécanismes de sélection pour les risques concurrents présentés dans [4].

Le lecteur pourra trouver dans [5], pour chacun des cas, avec et sans événement terminal, un exemple d'école pour illustrer la procédure ainsi que des simulations pour des modèles simples et solubles analytiquement qui permettent d'apprécier la qualité des estimateurs présentés. Nous illustrons ici nos méthodes sur deux essais thérapeutiques dans les sections 3.2.3 et 3.2.4. La section 3.2.5 propose des extensions des procédures proposées.

3.2.2 Modélisation

Soit $(T_{1p},\ p \in \mathbb{N}^*)$ (l'intérêt de l'indice 1 apparaîtra plus loin) une suite croissante de variables aléatoires strictement positives, de distributions $F_{1p}(t) = P(T_{1p} \leq t)$ et telles que $\forall q < p,\ F_{1p}(t) < F_{1q}(t)$. Ces variables représentent les délais d'occurrence d'un événement non terminal pour un individu, depuis une origine commune des temps ; p indice le rang d'un événement pour l'individu. A chaque T_{1p} peut être associé un processus de comptage défini par $N_p(t) = \mathbb{1}(T_{1p} \leq t)$, et $N(t) = \sum_{p=1}^{\infty} N_p(t)$ définit un processus global. $E\left(N(t)\right)$ représente l'espérance du nombre d'événements, tous rangs confondus, qui se produisent avant la date t dans une unité statistique. Elle correspond à une « incidence cumulée » totale au temps t. Notons que cette incidence peut être plus grande que 1. La propriété suivante est immédiate :

$$E\left(N(t)\right) = \sum_{p=1}^{\infty} F_{1p}(t). \tag{3.17}$$

Elle permet d'estimer $E\left(N(t)\right)$ dès lors qu'on dispose d'estimateurs pour les distributions $F_{1p}(t)$; cela s'applique en particulier aux estimateurs non paramétriques.

Un n-échantillon se présente sous la forme d'un ensemble de n séries indépendantes (individus) des variables $(T_{1p}, p \in \mathbb{N}^*)$. Chaque série ne comprend qu'un nombre fini de délais d'événements, variable d'une série à l'autre, du fait de la censure aléatoire à droite, qui vient masquer l'observation d'événements ultérieurs.

Il peut également arriver que la série des délais d'événements soit tronquée aléatoirement par la survenue d'un événement terminal, qui fait perdre définitivement au processus sa capacité de produire ultérieurement des récurrences. Ainsi, pour l'individu i, on a la série $T_{11i}, T_{12i}, \ldots, T_{1n_i i}$, à laquelle on doit ajouter un dernier délai : le délai de censure ou le délai jusqu'à l'événement terminal selon les cas, qui prend le rang $n_i + 1$.

Moyennant l'adjonction d'un modèle de censure et de troncature à droite, nous proposons des estimations non paramétriques convergentes des fonctions F_{1p} (ainsi que des distributions de censure et de troncature) et en conséquence une estimation de $E\left(N(t)\right)$ et des prévalences associées à un événement de rang donné.

3.2.3 Processus censuré sans événement terminal

Méthode d'estimation non paramétrique

Dans cette section, la production d'événements récurrents est supposée sans fin, mais nous considérons que, pour chaque individu, la série de délais se termine nécessairement par un délai de censure, au-delà duquel aucun événement n'est plus observable. L'individu i fournit donc n_i délais d'événements et un délai de censure, de rang $n_i + 1$.

Notre but étant d'obtenir des estimations non paramétriques convergentes des fonctions F_{1p}, fonctions de répartition des variables aléatoires $(T_{1p}, p \in \mathbb{N}^*)$, nous allons donc travailler à rang p fixé. La censure est unique pour chaque individu et vient masquer la série des événements récurrents à un rang donné, variable selon les individus. Son délai correspond à une variable aléatoire T_0. Au rang p, pour chaque individu de l'échantillon, ou bien la censure a eu lieu à un rang $k \leq p$ et le délai de l'événement récurrent T_{1p} est alors censuré par le délai de censure T_0, ou bien la censure est associée à un rang supérieur à p et le délai T_{1p} n'est pas censuré. Il paraît donc naturel d'imposer le mécanisme de censure suivant au rang p (notations classiques d'Efron) :

$$\left\{ \begin{array}{rl} Y_p & = \min(T_0, T_{1p}) \\ \Delta_p & = \mathbb{1}(T_{1p} \leq T_0) \end{array} \right. \tag{3.18}$$

où Y_p désigne le délai observé au rang p, et Δ_p est l'indicatrice d'événement ($\Delta_p = 1$ si on observe un événement et $\Delta_p = 0$ si l'on observe une censure). De plus, nous supposerons que les délais d'événements et de censure sont indépendants, quel que soit le rang associé à ces délais.

Les observations utiles pour l'estimation au rang p sont donc de deux sortes : les délais T_{1p} non censurés et les délais de censure T_0 associés à des rangs $k \leq p$. Notons qu'au rang p, chaque individu contribue une fois et une seule, si bien que l'on peut résumer les observations pertinentes au rang p par les couples $\{(Y_i, \Delta_i),\ i = 1, \ldots, n\}$ (pour ne pas alourdir les notations, l'indice p sera omis lorsque l'indice i intervient). Introduisons les notations suivantes au rang p :

$$\begin{array}{rcl} U_p(t) & = & P(Y_p \leq t), \\ U_{0p}(t) & = & P(Y_p \leq t, \Delta_p = 0), \\ U_{1p}(t) & = & P(Y_p \leq t, \Delta_p = 1). \end{array}$$

On a d'autre part :

$$\begin{array}{rcl} F_{1p}(t) & = & P(Y_p \leq t \mid \Delta_p = 1), \\ F_{0p}(t) & = & P(Y_p \leq t \mid \Delta_p = 0). \end{array}$$

Les fonctions U_p, U_{0p} et U_{1p} sont facilement estimables empiriquement ; F_{1p} et F_{0p} sont les fonctions d'intérêt à estimer. L'indépendance des processus d'événements et de censure ainsi que le mécanisme de censure (3.18) impliquent alors :

$$P(Y_p > t) = P(T_0 > t)\ P(T_{1p} > t), \tag{3.19}$$

c'est-à-dire

$$1 - U_p(t) \quad = \quad (1 - F_{0p}(t))\ (1 - F_{1p}(t)). \tag{3.20}$$

En notant R la variable aléatoire correspondant au rang associé à un délai de censure, on obtient la décomposition suivante de la fonction F_{0p} :

$$F_{0p}(t) = \sum_{k=1}^{p} P(Y_k \leq t, R = k \mid \Delta_k = 0) = \sum_{k=1}^{p} I_{0k}(t). \tag{3.21}$$

Cette décomposition nous suggère d'utiliser, par commodité d'expression et pour pouvoir replacer notre problème d'estimation dans un cadre déjà étudié, la notion de mélange : les délais de censure qui interviennent au rang p peuvent être considérés comme issus d'un mélange de délais de censure associés à des rangs $k \leq p$.

La relation (3.20) peut s'écrire :

$$1 - U_{0p}(t) - U_{1p}(t) = (1 - \sum_{k=1}^{p} I_{0k}(t))\ (1 - F_{1p}(t)). \tag{3.22}$$

Elle correspond à la relation 3.11 de [4], caractéristique du « mécanisme de mélange censuré », introduite pour des événements concurrents en présence d'une censure aléatoire à droite. Les risques concurrents peuvent en effet être modélisés en considérant un mélange d'événements dont l'observation peut être masquée du fait de la censure à droite. La situation est simplement inversée dans le présent article puisque le mélange est un mélange de censures de rang $k \leq p$, censuré par un événement, mais une démarche similaire peut s'appliquer. En conséquence, les estimations des quantités qui nous intéressent, à savoir les fonctions F_{1p} et F_{0p}, découlent d'un algorithme simple présenté dans [4].

Comme nous l'avons montré dans [4], l'estimateur de F_{1p} ainsi obtenu est l'estimateur de Kaplan-Meier au rang p, tout individu censuré à un rang inférieur ou égal à p étant considéré comme censuré pour le rang p :

$$\widehat{F}_{1p}(t) = 1 - \prod_{i:\delta_i=1} \left(1 - \frac{\mathbb{1}_{[y_i \leq t]}}{n - i + 1}\right). \tag{3.23}$$

Les estimations des fonctions $F_{0p}(t) = \sum_{k=1}^{p} I_{0k}(t)$, lois de mélanges de délais de censures de rangs inférieurs ou égaux à p, ne correspondent pas à des distributions d'intérêt. En revanche, F_{0r} correspond à la fonction de répartition de la censure, que nous noterons plus simplement F_0. En effet, au rang r, rang maximal observé, il n'y a plus que des délais de censure, si bien que la relation caractéristique (3.20) s'écrit :

$$1 - U_{0r}(t) = 1 - F_{0r}(t), \tag{3.24}$$

et donc l'estimateur empirique de U_{0r} constitue un estimateur non paramétrique de F_0, puisque la censure est entièrement observable : dans le contexte d'événements répétés, les événements n'empêchent pas la censure de se produire.

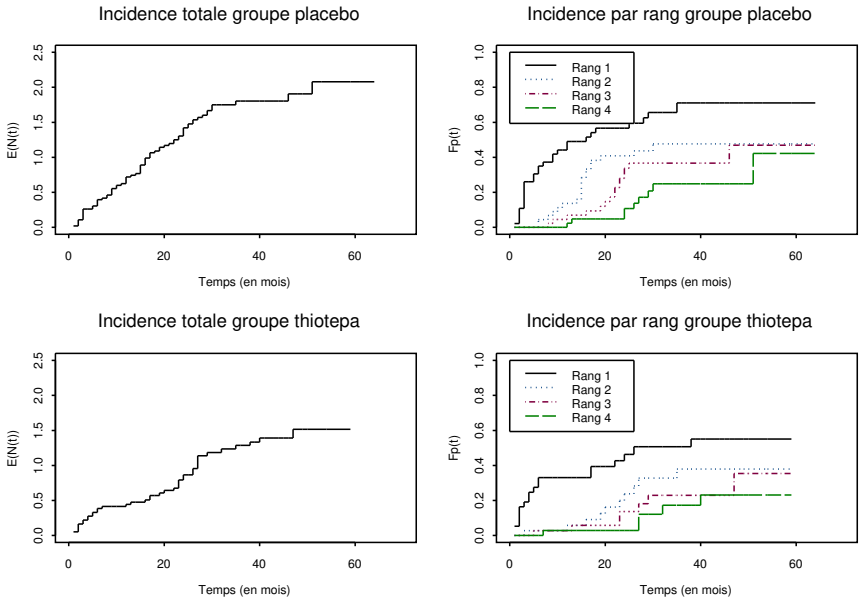


FIGURE 3.2 – Estimations des incidences cumulées totales et par rang pour l'essai `bladder`.

Un estimateur non paramétrique de l'incidence cumulée totale $E(N(t))$, espérance du nombre total d'événements répétés qui ont eu lieu avant t tous rangs confondus, s'obtient par sommation sur p (jusqu'au rang maximal $r-1$ où on observe des événements) des estimateurs non paramétriques des incidences $F_p(t)$ à chaque rang (cf (3.17)) :

$$\widehat{E(N(t))} = \sum_{p=1}^{r-1} \widehat{F}_{1p}(t). \tag{3.25}$$

Comme les estimateurs à chaque rang p sont convergents (en tant qu'estimateurs de Kaplan-Meier), leur somme l'est également.

Les prévalences dans un état p sont également estimables non paramétriquement à partir des estimations non paramétriques des fonctions $F_{1p}(t)$.

Application biomédicale : l'essai thérapeutique `bladder`

Nous avons repris la classique étude des cancers de la vessie de BYAR 1980 dont les données (`bladder`) sont disponibles en exemple dans le logiciel `R`, en considérant les deux bras placebo et thiotepa. Quand ils entrent dans l'essai, les 86 patients présentent des tumeurs de la vessie superficielles qui leur sont enlevées, puis ils sont randomisés dans les deux groupes de chimiothérapie (placebo ou thiotepa). Beaucoup de patients présentent des récurrences multiples de tumeurs au cours de l'étude, qui leur sont enlevées à chaque visite : 47 patients ont présenté une seule récurrence, 29 deux récurrences, 22 trois récurrences et 14 patients trois récurrences ou plus. Très peu de patients ont eu plus de 4 récurrences, donc seules les 4 premières récurrences ont été considérées.

Les estimations des incidences cumulées rang par rang et totale sont fournies dans la figure 3.2. On peut noter que, pour les 4 récurrences, les courbes des patients traités sont plus basses que celles des patients non traités.

3.2.4 Processus censuré tronqué par un événement terminal

Méthode d'estimation non paramétrique

Nous considérons à présent que les séries des délais des individus se terminent nécessairement soit par un *délai de perte* (la perte désigne l'événement terminal), au-delà duquel aucun événement ne peut plus se produire, soit par un *délai de censure*, au-delà duquel aucun événement ou perte n'est plus observable ; nous attacherons à ces délais, qu'il s'agisse d'une perte ou d'une censure, un rang, celui du dernier événement d'intérêt observé augmenté d'une unité. L'individu i fournit donc n_i délais d'événements et un délai de censure ou un délai de perte, de rang $n_i + 1$.

Notre but est d'obtenir des estimations non paramétriques convergentes des incidences cumulées des événements rang par rang, en présence de l'événement terminal. Nous allons donc travailler à rang p fixé. En plus du délai de censure T_0 existe maintenant un délai de perte T_2. La censure est unique pour chaque individu et vient masquer la série des événements récurrents à un rang donné, variable selon les individus, à moins que ce ne soit le délai de perte qui intervienne, en venant tronquer la série des événements récurrents à un rang donné, variable lui aussi selon les individus. Au rang p, pour chaque individu de l'échantillon, ou bien la censure a eu lieu à un rang $k \leq p$ et le délai de l'événement récurrent T_{1p} est alors censuré par le délai de censure T_0, ou bien la perte a eu lieu à un rang $k \leq p$ et le délai de l'événement récurrent T_{1p} est alors tronqué par le délai de perte T_2, ou bien la censure ou la perte, dont l'une seule des deux peut être observée, est associée à un rang supérieur à p et le délai T_{1p} n'est ni censuré ni tronqué.

Nous allons supposer, parce que cette hypothèse paraît raisonnable dans beaucoup d'études, que les délais de survenue des événements ainsi que les délais de survenue des pertes sont indépendants du délai de censure. Par contre, il n'y a pas de raison de supposer l'indépendance du délai de perte et du délai jusqu'à l'événement d'intérêt. On peut donc modéliser par commodité d'expression la compétition qui se joue entre la perte et l'événement de rang p par un mélange (voir article [4]). La variable résultante sera notée T_{12p} et sa fonction de répartition F_{12p}. Ces remarques conduisent à considérer le mécanisme suivant de censure et de troncature au rang p :

$$\left\{\begin{array}{lcl} Y_p & = & \min(T_0,T_{12p}) \\ \Delta_p & = & 0 \text{ si } T_0 < T_{12p} \\ \Delta_p & = & 1 \text{ si } T_{12p} \leq T_0 \text{ et } T_{12p} = T_{1p} \\ \Delta_p & = & 2 \text{ si } T_{12p} \leq T_0 \text{ et } T_{12p} = T_2 \end{array}\right. \tag{3.26}$$

où Y_p désigne le délai observé au rang p, et Δ_p est le type de ce délai ($\Delta_p = 1$ si on observe un événement, $\Delta_p = 2$ si on observe une perte et $\Delta_p = 0$ si on observe une censure).

Les observations utiles pour l'estimation au rang p sont donc de trois sortes : les délais T_{1p} non censurés et non tronqués, les délais de perte T_2 associés à des rangs $k \leq p$ et les délais de censure T_0 associés à des rangs $k \leq p$. Notons qu'au rang p, chaque individu contribue une fois et une seule, si bien que l'on peut résumer les observations pertinentes au rang p par les couples $\{(Y_i,\Delta_i),\, i = 1,\ldots,n\}$ (l'indice p sera omis lorsque l'indice i intervient).

En plus des fonctions U_p, U_{0p}, U_{1p} et F_{0p} déjà définies dans le cas sans perte (voir section 3.2.3), nous avons besoin des fonctions suivantes, définies pour tout rang p compris

entre 1 et r, rang maximal observé dans l'échantillon :

$$\begin{aligned} U_{2p}(t) &= P(Y_p \leq t, \Delta_p = 2), \\ I_{1p}(t) &= P(Y_p \leq t, \Delta_p = 1 \mid \Delta_p = 1 \text{ ou } 2), \\ I_{2p}(t) &= P(Y_p \leq t, \Delta_p = 2 \mid \Delta_p = 1 \text{ ou } 2), \\ F_{12p}(t) &= P(Y_p \leq t \mid \Delta_p = 1 \text{ ou } 2) = I_{1p}(t) + I_{2p}(t). \end{aligned}$$

L'indépendance de T_{0p} et de T_{12p} pour tout p ainsi que le mécanisme de censure et de troncature (3.26) conduisent, pour les mêmes raisons qu'à la section 3.2.3, à

$$P(Y_p > t) = P(T_{12p} > t)\ P(T_0 > t), \tag{3.27}$$

c'est-à-dire, pour tout rang p, $1 \leq p \leq r$:

$$\begin{aligned} 1 - U_p(t) &= 1 - U_{0p}(t) - U_{1p}(t) - U_{2p}(t) \\ &= (1 - F_{0p}(t))\ (1 - F_{12p}(t)) \\ &= (1 - F_{0p}(t))\ (1 - I_{1p}(t) - I_{2p}(t)). \end{aligned}$$

En notant R la variable aléatoire correspondant au rang associé aux délais de censure ou de perte, la décomposition (3.21) de la fonction F_{0p} proposée dans la section sans perte est toujours valable.

Remarquons qu'une décomposition similaire au rang p s'applique à la fonction I_{2p} :

$$I_{2p}(t) = \sum_{k=1}^{p} P(Y_k \leq t, R = k, \Delta_k = 2 \mid \Delta_k = 1 \text{ ou } 2). \tag{3.28}$$

Les délais de perte qui interviennent au rang p peuvent donc être considérés comme issus d'un mélange de délais de perte associés à des rangs $k \leq p$.

Nous avons donc au rang p un mécanisme de mélange censuré : un mélange constitué de l'événement d'intérêt de rang p et de délais de pertes de rangs $\leq p$ est censuré par minimum avec un mélange de délais de censures de rangs $\leq p$. Nous pouvons alors appliquer l'algorithme correspondant à ce cas proposé dans [4] et montrer facilement que l'estimateur obtenu pour F_{12p} est l'estimateur de Kaplan-Meier :

$$\widehat{F}^{\mathrm{KM}}_{12p}(t) = 1 - \prod_{i:\delta_i \in \{1,2\}} \left(1 - \frac{\mathbb{1}_{[y_i \leq t]}}{n - i + 1}\right). \tag{3.29}$$

L'estimateur obtenu pour I_{1p} est l'estimateur de Kalbfleisch et Prentice 1980 :

$$\widehat{I}^{Pr}_{1p}(t) = \sum_{i:\delta_i = 1} \frac{\mathbb{1}_{[y_i \leq t]}}{n - i + 1}\ \left(1 - \widehat{F}^{\mathrm{KM}}_{12p}(y_i)\right).$$

Les estimateurs de F_{0r} et de I_{2r} constituent des estimateurs non paramétriques de F_0, fonction de répartition de la censure et de F_2, fonction de répartition de la perte. En effet, il est aisé de constater qu'au rang r, il ne peut plus y avoir d'événement récurrent, et donc la relation caractéristique (3.28) s'écrit :

$$\begin{aligned} 1 - U_{0r}(t) - U_{2r}(t) &= (1 - F_{0r}(t))\ (1 - I_{2r}(t)) \\ &= (1 - F_0(t))\ (1 - F_2(t)). \end{aligned}$$

On voit donc que la censure et la perte se censurent mutuellement, et la procédure d'estimation de F_0 et de F_2 est donc la procédure habituelle de Kaplan-Meier, appliquée à l'échantillon de l'ensemble des pertes et des censures :

$$\begin{aligned}\widehat{F}_0^{\mathrm{KM}}(t) &= 1 - \prod_{i:\delta_i=0}\left(1 - \frac{\mathbb{1}_{[y_i \le t]}}{n - \ell(i,r) + 1}\right),\\ \widehat{F}_2^{\mathrm{KM}}(t) &= 1 - \prod_{i:\delta_i=2}\left(1 - \frac{\mathbb{1}_{[y_i \le t]}}{n - \ell(i,r) + 1}\right),\end{aligned}$$

où $\ell(i,r)$ désigne le rang de l'observation i dans l'échantillon constitué de l'ensemble des pertes et des censures à tous les rangs.

Un estimateur de l'incidence cumulée totale des événements d'intérêt, tous rangs confondus, s'obtient en sommant les estimateurs des incidences à chaque rang p :

$$\widehat{E(N(t))} = \sum_{p=1}^{r-1} \widehat{I}_{1p}(t).$$

Pour ce qui est de la prévalence associée aux événements de rang p, en présence d'événements terminaux, les individus qui sont entrés dans l'état p avec une probabilité $I_{1p}(t)$ peuvent à l'instant t soit produire un $(p+1)$-ème événement récurrent, avec la probabilité $I_{1,p+1}(t)$ qui les fait passer dans l'état suivant $p+1$, soit produire un événement terminal de rang $p+1$ avec la probabilité $P(Y \le t, R = p+1, \Delta = 2 \mid \Delta = 1 \text{ ou } 2) = I_{2,p+1}(t) - I_{2p}(t)$. D'où :

$$\begin{aligned}\pi_p(t) &= I_{1p}(t) - I_{1,p+1}(t) - (I_{2,p+1}(t) - I_{2,p}(t))\\ &= F_{12p}(t) - F_{12,p+1}(t), \quad 0 < p < r \qquad (3.30)\\ \pi_0(t) &= 1 - F_{12,1}(t).\end{aligned}$$

Application biomédicale : étude `CAPRIE`

Nous ré-analysons l'étude `CAPRIE` (Clopidogrel versus Aspirine in Patients at Risk of Ischaemic Events) du `CAPRIE` Steering Committee (Gent et al. 1996), un essai thérapeutique destiné à montrer la supériorité du clopidogrel sur l'aspirine dans la prévention des événements ischémiques. Environ 20 000 patients ont été suivis de 1 à 3 ans, et jusqu'à 3 récurrences ont été observées. L'analyse selon le protocole a porté sur un critère composite, à savoir le premier événement ischémique, fatal ou non. Les courbes de survie par groupe de traitement pour ce critère ont été estimées par la méthode de Kaplan-Meier et on obtient à l'aide du test du logrank une différence significative, qui atteint en différence relative 8 % à 3 ans en faveur du clopidogrel.

Nous reprenons cette étude en considérant tous les événements récurrents. Nous présentons à la figure 3.3 les courbes d'incidence cumulée par groupe des patients de la strate PAD (Peripheral Arterial Disease) ; il s'agit des patients inclus dans l'étude avec une artériopathie des membres inférieurs, qui les prédispose à produire des événements cardio-vasculaires (CV) graves, dont des infarctus du myocarde, des infarctus cérébraux ou encore à devoir être amputés du membre inférieur atteint par la maladie ; tous ces événements peuvent être éventuellement fatals. Dans cet exemple sont considérés comme événements récurrents les événements cardio-vasculaires non fatals, les événements cardio-vasculaires fatals constituant quant à eux des événements terminaux que nous avons regroupé sous le terme « décès toutes causes ». Il y a environ 3300 patients dans chaque groupe. On constate que la supériorité du clopidogrel sur l'aspirine

s'exprime à la fois vis-à-vis des événements terminaux et des événements récurrents, globalement et rang par rang. Les patients ayant reçu le clopidogrel ont eu au plus deux récurrences ; quelques patients ayant reçu l'aspirine en ont eu trois. Nous avons appliqué aux mêmes données le modèle de Cox généralisé de Wei et al. 1989, que l'on note WLW. Seules les deux premières récurrences ont pu être considérées, puisque le groupe ayant reçu le clopidogrel a eu deux récurrences au plus. Le décès a été traité comme une strate à part, et les événements récurrents tronqués sont considérés comme censurés. Les risques relatifs estimés sont les suivants : 1,23 (p[3] = 0,03) pour la première récurrence, 2,02 (p = 0,02) pour la deuxième récurrence, et 1,17 (p = 0,14) pour le décès, toujours en faveur du clopidogrel. Nous constatons sur les courbes de la figure 3.3 (en haut à droite) que les estimations des incidences rang par rang obtenues sont assez proches de nos estimations non paramétriques. De même, pour le décès, les courbes sont très proches. En ce qui concerne l'incidence globale, les estimations tous rangs confondus ont été calculées comme ce qui est fait classiquement par la méthode WLW en considérant un modèle de Cox avec paramètre unique pour prendre en compte l'effet du traitement sur les deux récurrences. Les courbes correspondantes sont les plus basses de la figure 3.3 en haut à gauche. Elles sont en quelque sorte une moyenne des courbes des rangs 1 et 2. Or, comme le montre la formule (3.17), la façon correcte d'obtenir une courbe d'incidence globale est de sommer les courbes des différents rangs, ce que nous avons également fait en sommant les courbes des rangs 1 et 2 de la méthode WLW. Nous constatons alors sur les courbes de la figure 3.3 en haut à gauche que les estimations obtenues sont assez proches de nos estimations non paramétriques. Sur ce même graphe, correspondant à tous les événements cardio-vasculaires non fatals, nos estimations montrent cependant une tendance : l'écart entre les deux groupes se resserre après 450 jours. Cette différence précoce est gommée par le modèle WLW du fait de l'hypothèse de proportionnalité des risques. Il se peut cependant que la tendance révélée par notre méthode soit réelle et que l'évolution ultérieure moins différenciée des courbes provienne des arrêts de traitement : en effet, les patients arrêtant le clopidogrel sont mis sous aspirine, mais l'analyse est faite en intention de traiter[4]. Cela mériterait une analyse plus fine. Nous avons également appliqué la méthode de Cook et Lawless 1997 de calcul de l'incidence globale aux données `CAPRIE`. Comme pour les simulations, les courbes d'incidence globale obtenues sont extrêmement proches des nôtres, à un point tel que nous ne les avons pas représentées. Rappelons que cette dernière méthode ne comporte pas d'estimation des incidences rang par rang, et exclut donc le calcul d'estimation des prévalences.

3.2.5 Conclusion et perspectives

Un des intérêts d'avoir une méthode non paramétrique disponible est de permettre la validation de modèles plus restrictifs. En particulier, l'hypothèse de proportionnalité des risques des modèles de Cox généralisés peut être graphiquement validée en comparant les estimations des incidences par groupe de ces modèles à nos estimations non paramétriques. Par exemple, les données `CAPRIE` semblent assez bien se prêter à une analyse par un modèle de risques proportionnels.

La méthode d'estimation proposée s'étend de façon immédiate à tout mélange d'événements récurrents de différents types, permettant l'estimation non paramétrique de

3. désigne le degré de signification du test de Wald testant la nullité du paramètre mesurant l'effet du traitement.

4. On parle d'analyse en intention de traiter lorsque des patients ayant abandonné le traitement que leur a affecté la randomisation restent considérés pour l'analyse dans leur groupe de traitement d'origine, quels que soient les traitements réellement administrés par la suite. C'est ce qui est fait habituellement dans les études cliniques de phase III.

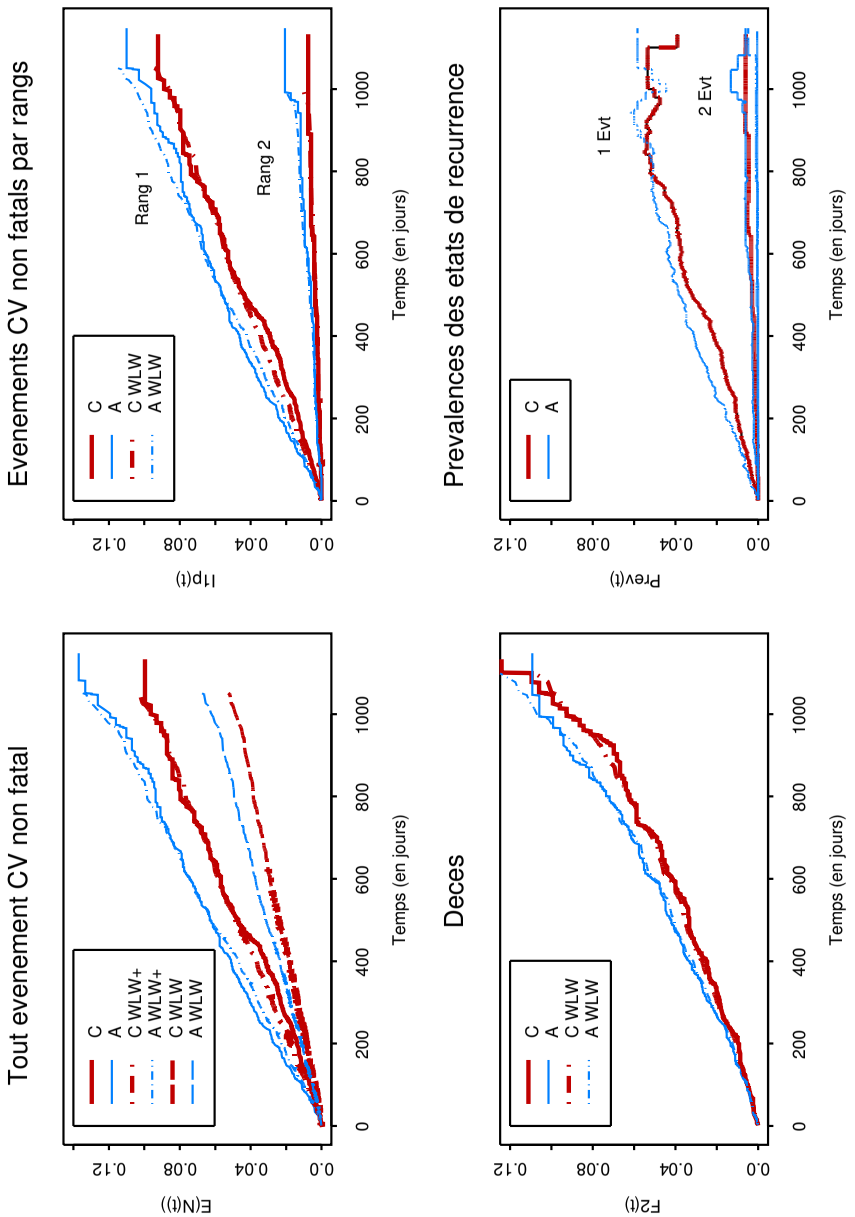


FIGURE 3.3 – Estimations des incidences cumulées et des prévalences pour l'étude CAPRIE (A = aspirine, C = clopidogrel ; les courbes WLW+ sont obtenues en sommant les courbes WLW des rangs 1 et 2).

l'incidence d'un type particulier d'événement. Si les événements terminaux sont également catégorisés, on sait que le modèle de mélange censuré permet d'estimer l'incidence d'une de leurs catégories. Cela permet par exemple d'estimer l'incidence des infarctus mortels ou non mortels et récurrents, en présence d'autres événements mortels ou non mortels et récurrents (par exemple des hospitalisations ou des décès pour une cause non cardio-vasculaire).

3.3 Suivi post-thérapeutique en oncologie

Dans cette section sont présentés les cinq articles écrits dans le cadre de l'encadrement de la thèse de Serge Somda, en codirection avec Thomas Filleron de l'Institut Claudius Regaud, dont le sujet était l'individualisation du suivi post-thérapeutique des patients traités du cancer en fonction du cancer et du type de rechute. La section 3.3.1 présente le contexte de la surveillance post-thérapeutique en oncologie. La section 3.3.2 présente l'état de l'art sur l'estimation de la durée de surveillance optimale post-thérapeutique avec en particulier l'approche de Mould et al. 2004. La section 3.3.3, qui correspond à l'article [13], présente une application de la méthodologie de Mould et al. 2004, développée initialement dans le cadre du cancer du sein, pour un site différent du cancer : il s'agit de déterminer la durée optimale de suivi de patients traités d'une tumeur germinale des testicules à un stade avancé, en tenant compte du pronostic et de la réponse au traitement du patient. Dans l'article [9], présenté à la section 3.3.4, nous généralisons l'approche de Mould et al. 2004, qui ne considère qu'un seul type d'événements, en prenant en compte plusieurs événements concurrents avec différentes probabilités de guérison associées. La section 3.3.5 résume l'article [12] qui compare la technique de l'article [9] pour déterminer la durée de surveillance avec une approche plus simple basée sur les risques proposée par Stewart-Merrill, Boorjian et al. 2015. Lorsque la durée de suivi ainsi que le nombre de visites ont été arrêtés par le médecin, la méthode proposée dans l'article [11] et décrite dans la section 3.3.6 permet de déterminer les dates optimales auxquelles les visites devraient être programmées. Nous proposons pour finir dans la section 3.3.7 un algorithme pour comparer les différentes stratégies de surveillance post-thérapeutique, qui correspond à la prépublication [23].

3.3.1 Motivation

Rappelons que le cancer est, d'après l'Organisation Mondiale de la Santé (who 2015), la première cause de décès dans le monde. En effet, 8,8 millions de décès dus au cancer ont été enregistrés en 2012 (Steward et Wild 2014) dont les principales localisations étaient, dans l'ordre du nombre de décès, le poumon, le foie, l'estomac, le cancer colorectal, le sein et l'œsophage.

Les patients en rémission à la suite d'un traitement curatif du cancer entrent dans la phase de surveillance post-thérapeutique. Cette phase vise à déceler les rechutes à un stade précoce, afin de proposer des traitements, curatifs ou palliatifs. La surveillance repose sur des examens cliniques, de l'imagerie, éventuellement des endoscopies et des études de marqueurs biologiques.

Idéalement, la phase de surveillance devrait durer jusqu'à ce que le patient soit considéré comme guéri, c'est-à-dire ne présentant plus de risque de rechute. Quant aux visites, elles devraient être les plus nombreuses possibles. Ceci n'est bien évidemment pas possible pour des raisons pratiques et de coût. Il est alors nécessaire de déterminer un calendrier de surveillance. Ce calendrier devra proposer des visites suffisamment fréquentes mais aussi en nombre suffisamment réduit afin de limiter la charge pour le patient, le praticien et l'établissement de soins.

Plusieurs instances spécialisées comme la Société Américaine d'Oncologie Clinique et la Société Européenne d'Oncologie Médicale ont proposé différentes recommandations pour le suivi et le calendrier de surveillance selon la pathologie. Par exemple, on trouve des recommandations pour le cancer du sein (Khatcheressian et al. 2006 ; Aebi et al. 2010 ; Cardoso et al. 2010), le poumon (Sorensen et al. 2010), les sarcomes (Casali et al. 2010), le cancer colorectal (Desch et al. 2005 ; Balmaña et al. 2010), etc. Ces différentes planifications, qui peuvent différer selon les pays, sont généralement définies sur la base d'avis d'experts, sans évaluation reposant sur des essais cliniques randomisés. En particulier, elles ne prennent pas en compte les caractéristiques propres au patient comme les facteurs pronostiques de rechute. Tous les patients sont en effet suivis de la même manière alors même que leurs délais de récidive du cancer peuvent différer en fonction de plusieurs facteurs individuels.

Étant donné le nombre croissant de patients en phase de suivi, les coûts liés à ce dernier sont en constante augmentation. Hofmann et al. 2002, après avoir analysé les données de 661 patients traités pour un mélanome en Allemagne entre 1983 et 1999, ont estimé que le coût moyen de la surveillance pour détecter une récidive avait augmenté de 5 806€ entre 1983 et 1987 et de 18 558€ entre 1987 et 1990, les surveillances les plus coûteuses étant celles pour lesquelles les risques de rechute sont les moins élevés. Il est alors évident que l'adaptation du suivi au risque de rechute du patient permettrait un contrôle plus important de ces coûts. Enfin, comme le patient en rémission est à risque de plusieurs types de récidives concurrents, la prise en compte de ces différents types d'événements dans la modélisation s'avère nécessaire.

3.3.2 Durée optimale du suivi post-thérapeutique : état de l'art

Les différentes recommandations sur la surveillance des patients en phase de rémission de leur cancer proposent généralement des calendriers de visites spécifiques pour les cinq premières années après le traitement (Hogendoorn et al. 2010 ; Balmaña et al. 2010 ; Cardoso et al. 2010). Au-delà de cinq ans, les calendriers conseillent le plus souvent d'organiser une visite par an, mais sans préciser pendant combien de temps devra durer la surveillance (Weaver et al. 2014). La planification de cette durée est pourtant cruciale dans la prévision des ressources financières et humaines.

La modélisation du délai d'apparition de la récidive ne peut s'appuyer sur les modèles classiques de survie qui considèrent que 100 % des individus réaliseront l'événement d'intérêt. Heureusement, une fraction des patients en rémission ne fera jamais de rechute, quelle que soit la durée de leur suivi et pourra être considérée comme guérie. Il faut donc utiliser des modèles de survie qui incluent cette probabilité de guérison (*cure rate models*).

Plusieurs auteurs ont proposé des méthodes de détermination du taux de guérison. Il peut être estimé par des techniques utilisant la survie relative (voir Dickman et al. 2004 ; Ederer et al. 1961). Tai et al. 2005 déterminent le nombre d'années de surveillance minimal pour pouvoir estimer le taux de guérison en modélisant la durée de survie par une distribution log-normale. Ils en concluent que cette durée de surveillance diffère énormément selon la localisation du cancer. Récemment, Ambrogi et al. 2014 ont utilisé une approche par les risques concurrents pour estimer le taux de guérison en comparant les mortalités spécifiques. Boag 1949 a proposé un modèle de survie intégrant directement un taux de guérison. Il s'agit d'un modèle de mélange paramétrique log-normal qui permet l'estimation à la fois de la distribution du délai jusqu'à la récidive et du taux de guérison. Mould et al. 2004, en utilisant le modèle de Boag, ont proposé une formule simple pour aider à la détermination de la durée de surveillance après traitement

pour des patientes ayant eu un cancer du sein détecté au stade précoce. Cette approche est présentée ci-dessous.

L'approche de Mould

Mould et al. 2004 proposent une règle de décision pour réduire la durée de surveillance post-thérapeutique dans le cadre du cancer du sein sans augmenter de façon significative la probabilité que des patientes rechutent après la fin du suivi alors qu'elles auraient pu être prises en charge avec succès. Leur objectif est la détection précoce de la majorité des récidives locales (c'est-à-dire dans le même sein) qui pourraient bénéficier avec succès d'un traitement curatif. Leur méthodologie est en deux étapes : le délai d'apparition de la la récidive locale (LR) est modélisé en utilisant le modèle de guérison de Boag, puis la durée de surveillance est déterminée en fonction de la distribution de ce délai.

Le modèle de guérison de Boag

Boag 1949 fait l'hypothèse qu'une proportion π $(0 \leq \pi \leq 1)$ de patients sera guérie (c'est-à-dire ne fera jamais de récidive) et qu'une proportion $(1-\pi)$ récidivera. Pour ce second groupe, le délai d'apparition de la récidive locale est modélisé par une loi log-normale. La fonction de survie sans récidive locale peut alors s'écrire comme un modèle de mélange :

$$S_{\text{LR}}(t) = \pi + (1-\pi)\left(1 - F^*_{\text{LR}}(t,\mu,\sigma)\right) \tag{3.31}$$

où F^*_{LR} est la fonction de répartition d'une distribution log-normale de moyenne μ et d'écart-type σ. F^*_{LR} étant une fonction de répartition, il s'ensuit que $\lim\limits_{t\to+\infty} F^*_{\text{LR}}(t) = 1$ et par conséquent $\lim\limits_{t\to+\infty} S_{\text{LR}}(t) = \pi$. La droite horizontale d'équation $y = \pi$ est donc une asymptote horizontale à la courbe de survie. Cette quantité représente le taux de guérison après le traitement. La figure 3.4 présente un exemple de fonction de survie du modèle de Boag pour un taux de guérison $\pi = 0{,}25$.

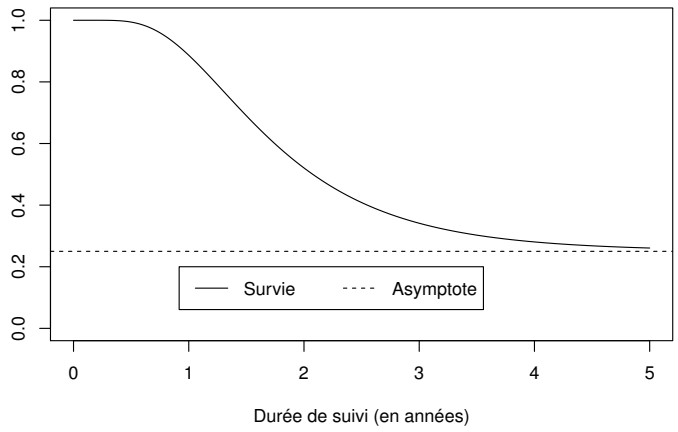


Figure 3.4 – Fonction de survie du modèle de Boag de paramètres $\mu = 3$, $\sigma = 0{,}5$ et $\pi = 0{,}25$.

Dans le cas où la distribution log-normale n'ajuste pas correctement les données, des variantes du modèle de Boag ont été proposées dans la littérature. A la place de la distribution log-normale, Sposto 2002 propose d'utiliser une distribution de Weibull.

Gamel et Vogel 1997 reprennent le modèle de mélange en incorporant une distribution log-logistique ou une distribution de Weibull. Peng, Dear et al. 1998 proposent une famille généralisée de modèles à laquelle on peut associer des effets aléatoires (voir aussi Peng et J. Zhang 2008 ; Peng et Taylor 2010). La distribution des événements peut également dépendre de facteurs pronostiques (Sposto 2002 ; Peng et J. Zhang 2008 ; Peng et Taylor 2010). Des procédures de type em (*Expectation-Maximization*) ont été développées pour l'estimation des paramètres de ces modèles (Peng et J. Zhang 2008 ; J. J. Zhang et M. Wang 2009 ; Lai et Yau 2009 ; Peng et Taylor 2010) et elles sont implémentées dans les logiciels statistiques les plus courants (Corbière et Joly 2007 ; Lambert 2007 ; Cai et al. 2012).

Détermination de la durée de surveillance

La deuxième étape de l'approche de Mould est la détermination de la durée de suivi. Les auteurs font l'hypothèse que toutes les lr sont détectées et notent ν $(0 \leq \nu \leq 1)$ la probabilité d'être traité avec succès en cas de récidive. La population qui entre dans la phase de surveillance post-thérapeutique peut donc se décomposer en trois sous-groupes (voir diagramme 3.5) :

1. une proportion π de patients qui ne vont jamais récidiver,
2. une proportion $\nu(1-\pi)$ de patients qui vont récidiver et dont la lr sera traitée avec succès,
3. une proportion $(1-\nu)(1-\pi)$ de patients qui vont récidiver mais être traités sans succès.

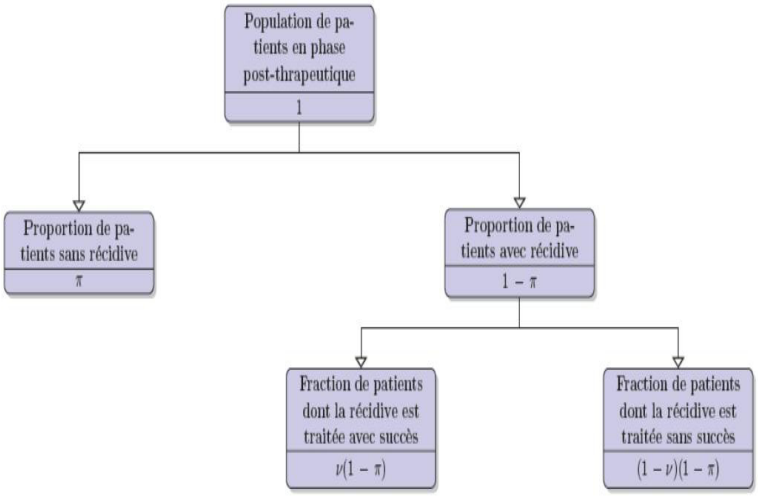


Figure 3.5 – Diagramme des différentes issues du traitement (Mould et al. 2004).

Mould et al. 2004 proposent de réduire la durée de la surveillance sans réduire de façon significative la proportion de patients du deuxième sous-groupe. La durée optimale du suivi, notée t_{opt}, est déterminée sous la contrainte que la proportion de patients dont la lr aurait été détectée après la fin du suivi et qui auraient été traités avec succès soit inférieure à une valeur fixée ε. On a donc

$$\nu(1-\pi)\left(1-F^{*}_{\mathrm{LR}}(t_{opt},\mu,\sigma)\right) \leq \varepsilon, \tag{3.32}$$

soit

$$\nu\left(S_{\mathrm{LR}}(t_{opt},\pi,\mu,\sigma)-\pi\right) \leq \varepsilon. \tag{3.33}$$

Si N patients entrent dans la phase de surveillance post-thérapeutique, le nombre attendu de récidives qui auraient pu être traitées avec succès après la fin du suivi vaut :

$$N_{\text{LR}}(t_{opt}) = N\nu\left(S_{LR}(t_{opt},\pi,\mu,\sigma) - \pi\right).$$

Mould et al. 2004 ont appliqué leur méthodologie à une population de femmes traitées pour un cancer du sein entre 1981 et 1990 et suivies pendant la durée recommandée de 10 ans. Par le modèle de Boag, pour $\pi = 0{,}85$, la fonction de survie sans récidive dans la population susceptible de rechuter est estimée à 10 % à 4 ans (soit $F^*_{\text{LR}}(4) = 0{,}90$). Avec leur approche, en posant $\nu = 0{,}80$, si le suivi n'avait duré que 4 ans au lieu de 10 ans, la proportion de patientes qui auraient rechuté après la surveillance et qui auraient été traitées avec succès est estimée à 0,12 %. En effet :

$$\begin{aligned}\nu \times [S_{\text{LR}}(4) - \pi] = &\ \nu \times [(\pi + (1-\pi)(1 - F^*_{\text{LR}}(4)) - \pi] \\ = &\ 0{,}80 \times [(0{,}85 + (1 - 0{,}85) \times 0{,}10) - 0{,}85] \\ = &\ 0{,}012.\end{aligned}$$

Pour $N = 1000$ patientes rentrant en phase de surveillance, cela correspond donc à $N_{\text{LR}}(4) = 12$. La durée de surveillance peut donc être réduite à 4 ans au lieu de 10 ans, en tolérant la "perte" de 12 patientes parmi 1000, patientes dont la récidive aurait pu être traitée avec succès si elles avaient été suivies pendant leur vie entière. Pour 8 ans de suivi, les mêmes calculs donnent $N_{\text{LR}}(8) = 6$.

3.3.3 Application au cancer des testicules

Dans l'article [13], nous avons appliqué l'approche de Mould au cancer des testicules pour présenter cette méthode aux oncologues spécialistes de ce domaine. Les tumeurs germinales du testicule représentent environ 1% de tous les cancers nouvellement diagnostiqués chez l'homme et ont causé 10 000 décès en 2012 (Ferlay et al. 2013). Le cancer du testicule est un cancer de bon pronostic, y compris en situation métastatique, avec une survie relative à 5 ans de 98-99 % pour les formes localisées et supérieure à 70 % pour les formes métastatiques. Les tumeurs germinales non séminomateuses (TGNS) correspondent à 40% des tumeurs germinales (Biggs et Schwartz 2007). L'article [13] donne des indications sur la détermination optimale de la durée de surveillance post-thérapeutique pour des patients traités pour des TGNS en phase métastatique. Les données utilisées sont issues de deux essais cliniques du Groupe Génito-Urinaire de la Fédération Française des Centres de Lutte Contre le Cancer pour les patients de bon pronostic (Essai GETUG T93BP, voir Culine, Kerbrat et al. 2007) et pour les patients de pronostic intermédiaire à mauvais (Essai GETUG T93MP, voir Culine, Kramar et al. 2008).

3.3.4 Durée de surveillance optimale en présence de risques concurrents

L'approche de Mould est un outil d'aide à la décision permettant la réduction de la durée de suivi. Elle a l'avantage d'être facile à comprendre par les praticiens. Cependant, cette approche ne permet de ne prendre en compte qu'un seul type de récidive. Or, en pratique, les patients en rémission de leur cancer sont la plupart du temps à risque de plusieurs types de récidives : récidives locales ou régionales, métastases à distance, second cancers primaires. De plus, les distributions associées à ces différents types de récidives ne sont pas identiques et ne sont pas forcément influencées par les mêmes facteurs pronostiques. Nous proposons dans l'article [9] une approche qui généralise celle de Mould et al. 2004 pour déterminer la durée optimale de suivi en prenant en compte plusieurs types concurrents de récidive, qui va être décrite ci-dessous.

Bien que les différents types de récidives envisagés ne soient pas exclusifs au sens propre du terme, comme nous nous intéressons uniquement au premier événement, la détection d'une récidive invalide l'observation des autres [5] et une approche par les risques concurrents est donc justifiée. Le modèle de guérison de Boag 1949 ne permet pas de modéliser la survie sans récidive en considérant plusieurs types de récidives concurrents. C'est donc le modèle paramétrique d'estimation en présence de risques concurrents de Jeong et Fine 2006 que nous avons utilisé et que nous explicitons ci-dessous après avoir défini nos notations.

Notations

Nous considérons une population à risque de K événements concurrents. T est le délai d'apparition de l'événement qui survient en premier depuis la fin du traitement. Δ est l'indicatrice d'événement : elle prend la valeur k ($k = 1,\ldots,K$) si l'événement de type k a été observé en premier pendant la durée de surveillance et la valeur 0 si aucune récidive n'a été observée pour l'individu pendant sa durée de surveillance (T est alors censuré à droite). Remarquons que cela correspond au mécanisme de sélection de mélange censuré 3.9 de [4]. Nous notons ν_k ($0 \leq \nu_k \leq 1$) pour $k = 1,\ldots,K$ la probabilité pour un patient d'être traité avec succès en cas de détection précoce d'un événement de type k.

La figure 3.6 généralise le diagramme de la figure 3.5 en considérant trois types de récidives pour le cancer du sein (récidive loco-régionale, cancer controlatéral [6] et métastases à distance). L'objectif de la surveillance est de détecter le maximum de récidives dans les boîtes les plus foncées.

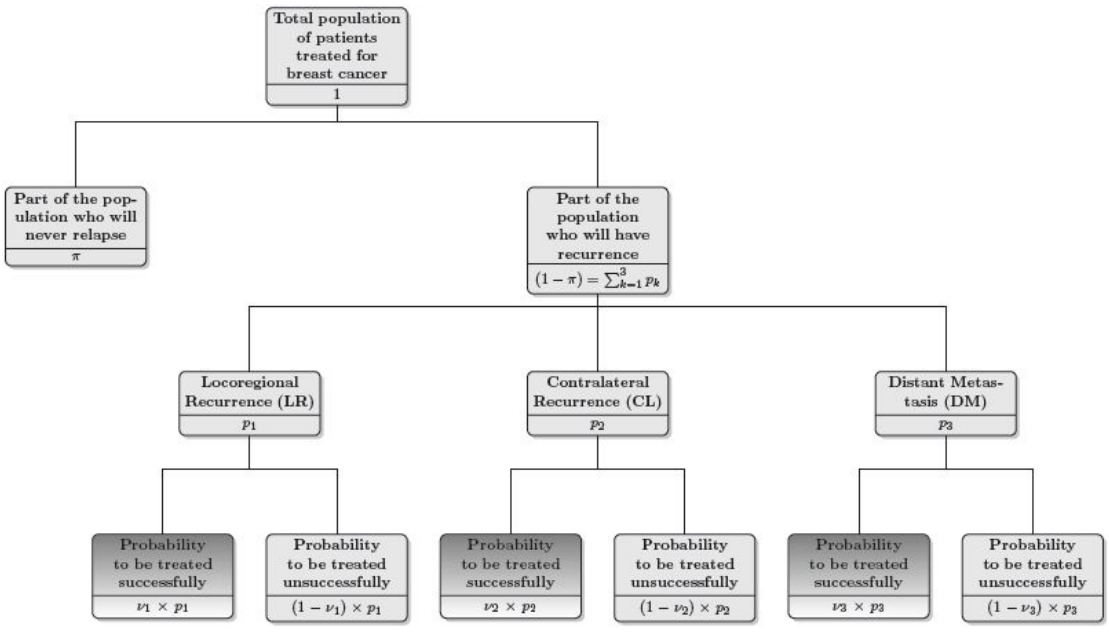


Figure 3.6 – Diagramme des différentes issues du traitement en considérant trois types concurrents de récidives et un taux de guérison π. p_k ($k = 1,2,3$) est la probabilité que l'événement de type k apparaisse en premier.

5. En fait, dès qu'un type de récidive est détecté, le patient retourne en phase de traitement et quitte la phase de surveillance post-thérapeutique : ainsi, l'apparition d'une récidive d'un autre type ne peut donc plus être observée en phase de suivi.

6. Il s'agit de l'apparition d'un cancer dans l'autre sein.

Nous nous intéressons aux fonctions d'incidences cumulées (notées en abrégé CIF) I_k, $k = 1,\ldots,K$, qui sont les probabilités que l'événement de type k survienne avant t en présence des autres événements :

$$I_k(t) = P(T \leq t,\Delta = k). \tag{3.34}$$

Ces fonctions sont des sous-distributions. Elles correspondent aux fonctions $I_{1,k}$ de la formule (3.10) de [4] ; l'indice 1 a été omis ici pour alléger les notations. Soit $S(t)$ la fonction de survie globale sans récidive. On a la relation suivante (identique à 3.10) :

$$S(t) = P(T > t) = 1 - \sum_{k=1}^{K} I_k(t).$$

La modélisation de Jeong et Fine avec taux de guérison

Forme générale du modèle

Jeong et Fine 2006 proposent une approche directe qui consiste à modéliser séparément les CIF associées à chaque type d'événement. Comme les CIF sont des sous-distributions, elles sont modélisées par des distributions impropres de Gompertz (Gompertz 1825), ce qui donne pour la CIF associée à l'événement de type k :

$$I_k(t,\alpha_k,\beta_k) = 1 - \exp\left\{\frac{\beta_k}{\alpha_k}(1 - e^{\alpha_k t})\right\} \text{ avec } \alpha_k < 0 \text{ et } \beta_k > 0. \tag{3.35}$$

Rappelons que le modèle de Gompertz est traditionnellement utilisé en analyse de la survie dans la biologie du vieillissement. Le paramètre de forme α_k est alors appelé *coefficient du taux de mortalité âge-dépendant* tandis que le paramètre d'échelle β_k est appelé *coefficient du taux de mortalité âge-indépendant* (voir Witten et Satzer 1992). Ces modèles sont aussi largement utilisés en démographie où ils permettent d'estimer les durées de vie des populations (Bagnoli et Bergstrom 2005).

On a

$$\lim_{t\to+\infty} I_k(t) = 1 - e^{\frac{\beta_k}{\alpha_k}} = p_k < 1$$

et la distribution de Gompertz est donc impropre. p_k est la probabilité que l'événement de type k soit observé en premier. Il s'ensuit que la proportion π d'individus qui ne feront jamais de récidive, qui correspond au taux de guérison, vaut

$$\pi = 1 - \sum_{k=1}^{K} p_k = 1 - K + \sum_{k=1}^{K} e^{\frac{\beta_k}{\alpha_k}}.$$

Introduction de covariables dans le modèle

Jeong et Fine 2007 ont proposé l'introduction de covariables dans le modèle précédent pour prendre en compte les facteurs pronostiques. Soit X le vecteur des covariables et γ_k le vecteur des paramètres de régression associés à l'événement de type k. Le modèle pour la CIF associée à l'événement de type k s'écrit :

$$I_k(t,\alpha_k,\beta_k,\tau_k,\gamma_k,X) = 1 - \left(1 - \tau_k \frac{\beta_k}{\alpha_k} e^{\gamma_k' X}(1 - e^{\alpha_k t})\right)^{-\frac{1}{\tau_k}} \tag{3.36}$$

où τ_k est un paramètre de flexibilité qui permet d'incorporer différents effets des covariables. Lorsque $\tau_k = 1$, on obtient un modèle à côtes proportionnelles. Lorsque $\tau_k \to 0$, le modèle 3.36 devient un modèle à risques proportionnels, qui s'écrit plus simplement

$$I_k^{\text{PH}}(t,\alpha_k,\beta_k,\gamma_k,X) = 1 - \exp\left(\frac{\beta_k}{\alpha_k} e^{\gamma_k' X}(1 - e^{\alpha_k t})\right). \tag{3.37}$$

C'est ce modèle que nous retiendrons dans toute la suite. Les paramètres α_k, β_k et γ_k $(k = 1, \ldots, K)$ peuvent être estimés par maximum de vraisemblance.

La fonction de perte

Le modèle précédent permet d'estimer la CIF pour chaque type de récidive à un temps donné. Cela permet de donner au clinicien l'estimation de la probabilité de rechuter après un délai de surveillance donné et si cette probabilité est en dessous d'un certain seuil, il peut alors décider de stopper la surveillance. Rappelons que les probabilités ν_k d'être traité avec succès en cas de détection d'une récidive varient selon le type de récidive. Cela justifie d'appliquer l'approche de Mould à chaque type d'événement séparément. On peut donc estimer une fonction de perte pour chaque type d'événement en pondérant la probabilité que cet événement se produise après la durée de surveillance par la probabilité de guérison associée. La fonction de perte globale, pour une durée de surveillance t, peut alors être définie comme la probabilité d'avoir une récidive après t qui aurait été traitée avec succès si elle avait été détectée.

La probabilité pour un patient d'avoir un événement de type k comme premier événement est $p_k = \lim_{t \to +\infty} I_k(t)$. Donc la probabilité d'avoir un événement de type k comme premier événement après t vaut $p_k - I_k(t)$. En tolérant un niveau de perte fixé ε comme dans MOULD et al. 2004, nous pouvons alors en déduire la durée optimale de suivi t_{opt}. La fonction de perte liée à l'événement de type k vaut $\varepsilon_k(t) = \nu_k \left(p_k - I_k(t)\right)$ et la fonction de perte globale s'en déduit simplement comme la somme des fonctions de perte par événement :

$$\varepsilon(t) = \sum_{k=1}^{K} \varepsilon_k(t) = \sum_{i=1}^{K} \nu_k \left(p_k - I_k(t)\right). \tag{3.38}$$

L'estimation des CIF I_k par le modèle de Jeong et Fine, avec ou sans covariables, permet donc d'estimer $\varepsilon(t)$, qui est la proportion de patients dont la récidive aurait pu être traitée avec succès si leur surveillance s'était prolongée indéfiniment après t. Cette méthode a été programmée avec le logiciel `R` par Serge Somda.

Illustration

Nous présentons une application de la méthode sur des données réelles de patients traités pour des sarcomes des tissus mous, qui désignent un ensemble de cancers localisés dans les tissus mous du corps (muscles, graisse, tissus fibreux). Les données utilisées proviennent du Groupe Sarcome de la Fédération Française des Centres de lutte contre le Cancer. La durée de surveillance des patients est en général de 10 ans mais les médecins adaptent l'intensité et les méthodes de suivi en fonction des facteurs de risque (voir BEITLER et al. 2000 ; GERRAND et al. 2007). Les patients traités pour un sarcome des tissus mous qui entrent en phase de surveillance post-thérapeutique sont à risque de trois types d'événements : une récidive locale, des métastases à distance ou le décès sans récidive. Comme l'objectif est de détecter le premier événement, nous considérons ces événements comme concurrents.

Description des données

Les données concernent 1614 patients âgés de 18 à 85 ans à la date du diagnostic de leur cancer (médiane des âges à 58 ans). 50,2 % étaient des hommes. En ce qui concerne le premier événement observé, 203 patients ont eu une récidive locale, 174 ont eu des métastases à distance et 21 sont décédés sans récidive.

	Récidive locale	Métastases à distance	Décès sans récidive	Total
α_k (s)	-9,31 (4,72)	-9,38 (5,08)	-2,08 (13,44)	
β_k (s)	5,32 (0,58)	4,54 (0,54)	0,42 (0,14)	
$I_k(5 \text{ ans})$ (s) (en %)	21,69 (1,59)	18,81 (1,51)	2,32 (0,59)	42,82
$I_k(\infty)$ (en %)	43,50	38,89	18,11	100

TABLE 3.1 – Résultats de l'estimation du modèle de Jeong et Fine sur les données des sarcomes des tissus mous. Les valeurs des paramètres et de leur écart-type estimé s ont été multipliées par 10^2. Les écarts-types estimés de α_k et β_k ont été obtenus par maximum de vraisemblance. Les écarts-types estimés des CIF à 5 ans ont été estimés à l'aide de la delta-méthode multivariée.

Estimation sans covariables

La première étape dans la détermination de la fonction de perte est l'estimation des paramètres des modèles de Jeong et Fine pour les trois CIF. Les résultats sont donnés dans le tableau 3.1. A titre de validation de ce modèle paramétrique, pour chacun des trois événements, nous avons également estimé non paramétriquement la CIF par l'estimateur de KALBFLEISCH et PRENTICE 1980 (voir formule 3.15) en utilisant le paquet `cmprsk` de `R` (voir GRAY 2011). La figure 3.7 **A** montre que les courbes obtenues pour les estimations paramétriques et non paramétriques des trois CIF sont très proches, ce qui valide le modèle de Jeong et Fine. La figure 3.7 **B** présente les fonctions de perte pour les trois événements ainsi que la fonction de perte globale, estimées en fixant des valeurs communément admises pour les probabilités ν_k de traiter avec succès les différents événements : $\nu_1 = 0{,}80$ pour la récidive locale, $\nu_2 = 0{,}05$ pour les métastases à distance et bien évidemment $\nu_3 = 0$ pour le décès. Il s'ensuit que la fonction de perte associée au décès est nulle et celle associée aux métastases à distance est très basse. Comme cet événement est de très mauvais pronostic, il n'est pas nécessaire de prolonger le suivi afin de le détecter. La figure 3.7 **C** présente la fonction de perte assortie d'un intervalle de confiance à 95 % : si par exemple la surveillance est stoppée au bout de 7 ans, 14,1 % des patients (IC : [5,1 ; 33,3]) seraient « perdus ». Il apparaît donc difficile de prendre ici une décision sur la durée optimale de surveillance. Cela sera plus simple en segmentant la population selon les valeurs des covariables.

Estimation avec covariables

Le facteur pronostique le plus important pour le sarcome des tissus mous est le grade histologique (COINDRE 2006). C'est une mesure à trois niveaux qui résume l'information pronostique (différentiation de la tumeur, index mitotique et importance de la nécrose). Nous possédons cette information pour 1407 patients (sur 1614). L'estimation des paramètres du modèle est donnée dans le tableau 3.2. Le niveau de référence choisi pour le grade histologique est le grade III qui est le grade de moins bon pronostic. Ainsi, le paramètre γ_{1k} (respectivement γ_{2k}) correspond au logarithme du risque relatif du groupe classé en grade I (respectivement grade II) par rapport au groupe de grade III pour l'événement de type k. Les courbes des CIF et les fonctions de perte par grade sont présentées à la figure 3.8. Ces fonctions de perte ont été calculées en gardant les mêmes valeurs que précédemment pour les paramètres ν_k. On constate que la fonction de perte globale pour les patients de grade I prend des valeurs faibles comparées à celles des patients de grade II et III. Bien qu'il y ait plus d'événements observés pour les patients de grade III que pour ceux de grade II, la fonction de perte pour les grade III n'est pas plus élevée. Cela est dû au fait que la plupart des événements observés chez les patients de grade

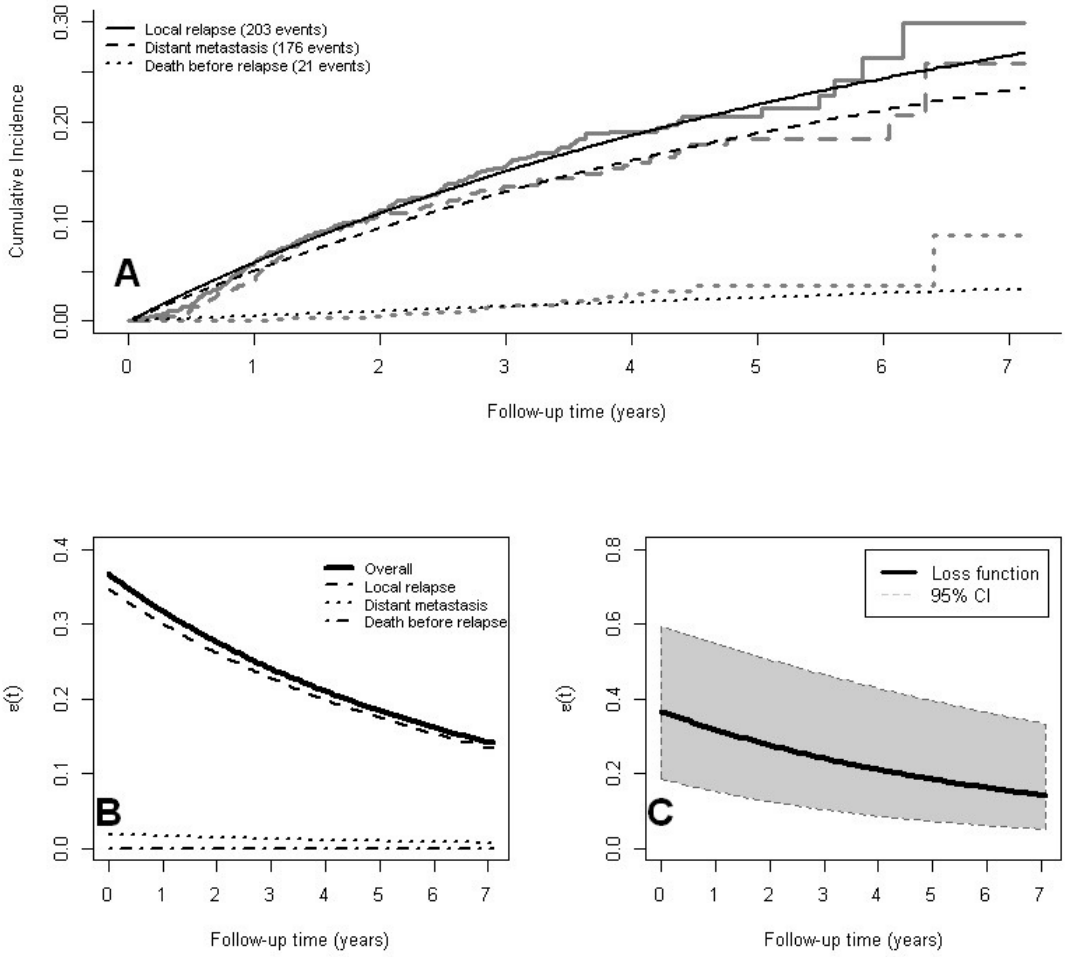


FIGURE 3.7 – **A** CIF estimées par le modèle paramétrique de Jeong et Fine (en noir) et par l'estimateur non paramétrique de Kalbleish et Prentice (en gris). **B** Fonctions de perte associées aux différents événements et globale. **C** Fonction de perte globale avec l'intervalle de confiance à 95 %.

	Récidive locale	Métastases à distance	Décès sans récidive
α_k ($\times 10^3$)	-11,73	-14,28	-3,61
β_k ($\times 10^3$)	5,41	9,70	0,53
γ_{1k}	-0,38	-2,23	-0,31
γ_{2k}	0,26	-0,83	-0,29
I_k(5 ans | Grade I) (en %)	14,75	4,13	3,13
I_k(5 ans | Grade II) (en %)	25,98	15,68	2,13
I_k(5 ans | Grade III) (en %)	20,78	32,35	2,84
I_k(10 ans | Grade I) (en %)	21,22	5,84	8,49
I_k(10 ans | Grade II) (en %)	36,21	21,57	3,81
I_k(10 ans | Grade III) (en %)	29,41	42,69	5,07

TABLE 3.2 – Résultats de l'estimation du modèle à risques proportionnels de Jeong et Fine sur les sarcomes des tissus mous avec la covariable Grade histologique (3 niveaux de gravité croissante ; le grade III est choisi comme référence).

III sont des métastases à distance, qui sont de très mauvais pronostic et qui n'ont donc que très peu d'influence sur la fonction de perte. A titre d'exemple, si le médecin fixe à 15 % le taux de perte acceptable, la surveillance des patients de grade I pourrait être stoppée à 28 mois (2 ans et 4 mois), alors que celle des patients de grade II et III serait plus longue (62 et 53 mois respectivement).

Discussion

La méthode proposée permet donc d'inclure plusieurs types de récidives dans la détermination de la durée optimale de suivi mais aussi d'individualiser la durée de suivi en fonction de groupes de patients définis par des facteurs pronostiques.

La distribution de Gompertz sur laquelle repose ce modèle impose que le risque de récidive soit décroissant avec le temps. Si ce n'est pas le cas, on peut envisager de la remplacer par des modèles de mélange paramétriques (voir LAU et al. 2011).

Une des limitations du modèle de Jeong et Fine est le fait de modéliser séparément les CIF des différents types d'événement. La conséquence est que, si le taux de guérison est nul, rien ne garantit que la somme des CIF convergera vers 1 à l'infini. Il faudrait envisager une estimation conjointe des CIF comme cela a été fait dans HUDGENS et al. 2011 mais cela serait sans doute plus lourd numériquement.

Enfin, on pourrait imaginer améliorer la fonction de perte en y incorporant des probabilités de succès du traitement en cas de récidive ν_k qui seraient alors estimées alors qu'elles ont été considérées comme connues dans notre méthode. En pratique, ces probabilités dépendent fortement des facteurs pronostiques et du délai de récidive depuis la fin du traitement.

3.3.5 Comparaison avec une approche basée sur les risques

Récemment, une approche très simple basée sur les risques a été proposée par STEWART-MERRILL, BOORJIAN et al. 2015 pour déterminer la durée de surveillance post-thérapeutique de patients traités pour un carcinome urothélial qui ont eu une cystectomie (ablation de la vessie). Les auteurs estiment, à l'aide de modèles paramétriques de Weibull, les risques de récidive et de décès d'autres causes pour ces patients et proposent de stopper la surveillance quand le risque de décès d'autres causes dépasse le risque de récidive après cystectomie. Cette approche avait aussi été proposée pour le

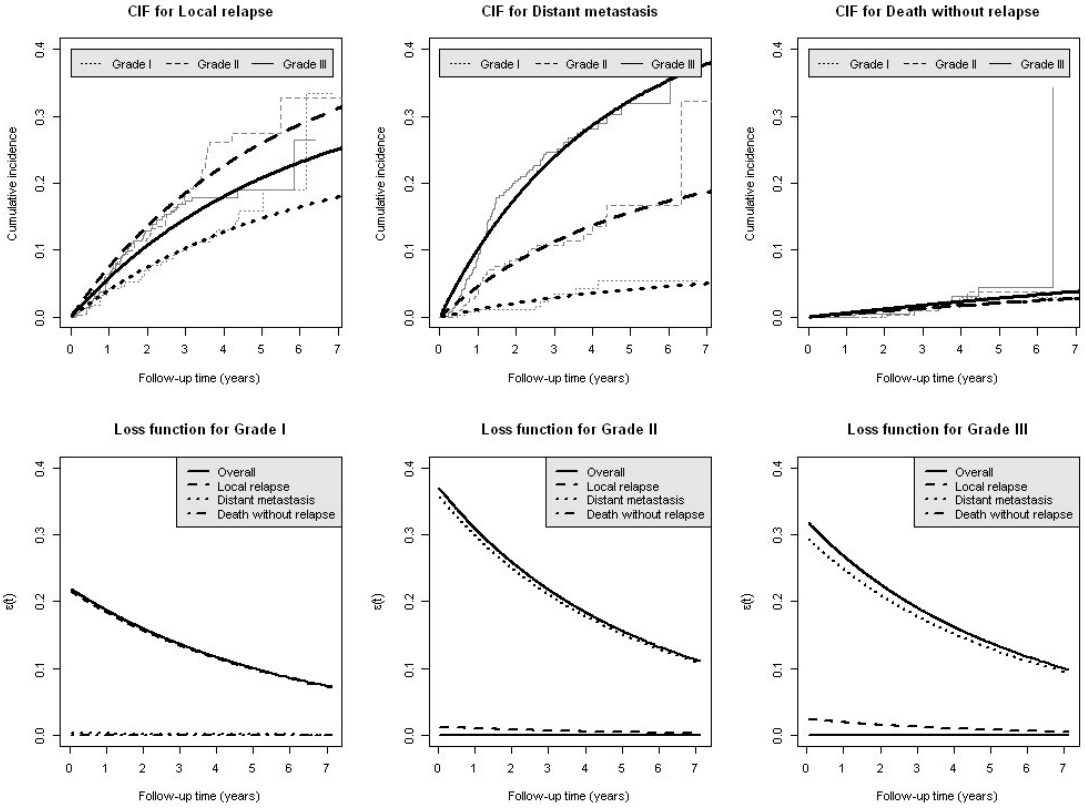


FIGURE 3.8 – **En haut :** CIF selon le grade histologique et le type d'événement estimées par le modèle paramétrique de Jeong et Fine (en noir) et l'estimateur non paramétrique de Kalbfleish et Prentice (en gris). **En bas :** fonctions de perte associées aux différents événements et globale selon le grade histologique.

suivi des carcinomes des cellules rénales (Stewart-Merrill, R. H. Thompson et al. 2015). Dans la lettre à l'éditeur [12], nous simulons trois scénarios très différents mais qui aboutissent au même délai de surveillance de 5 ans en se basant sur les estimations des risques de Stewart-Merrill, Boorjian et al. 2015. Nous appliquons notre méthode [9] de détermination de la durée optimale de suivi avec risques concurrents, basée sur l'estimation des CIF et nous déterminons les nombres de patients qui auraient pu être traités avec succès s'ils avaient été suivis après la durée de surveillance, pour les trois scénarios, différentes durées de surveillance et différentes valeurs des probabilités de guérison de la récidive. Nous montrons alors que nos préconisations en termes de durées de surveillance optimales sont très différentes selon les différents paramètres simulés alors que l'approche de Stewart-Merrill, Boorjian et al. 2015 basée sur les risques conseille une durée de surveillance unique de 5 ans dans tous les cas. En effet, considérer uniquement les risques de récidive et de décès ne donne pas d'information sur la proportion des patients qui récidivent pendant une période donnée. De plus, un événement concurrent comme le décès d'autres causes peut se produire avant la récidive, ce qui n'est pas pris en compte dans les estimations de Stewart-Merrill, Boorjian et al. 2015. Enfin, la durée de surveillance doit être adaptée en tenant compte du niveau du risque de récidive et doit tenir compte des probabilités de succès du traitement en cas de détection précoce.

3.3.6 Planification optimale des visites de contrôle

Une fois la durée optimale de surveillance déterminée se pose le problème de la planification optimale des visites de contrôle. C'est l'objet de l'article [11]. Beaucoup de travaux ont été publiés dans le cadre du dépistage du cancer pour proposer des calendriers optimaux pour une détection précoce du cancer (voir S. J. Lee et Zelen 1998 ; Shen et Zelen 2001 ; Shen et Parmigiani 2003 ; Shen et Zelen 2005 ; S. Y. Lee et al. 2007 ; Draisma et Van Rosmalen 2013). En effet, plus le cancer est détecté tôt et meilleures sont les chances de guérison. Par exemple, dans le cadre du cancer du sein, S. J. Lee et Zelen 1998 fournit un calendrier de dépistage basé sur la maximisation d'une fonction d'utilité proposée par Zelen 1993 : la probabilité de détection de la maladie à un stade préclinique (c'est-à-dire avant l'apparition de symptômes) est maximisée tandis que la probabilité de détection à un stade clinique plus tardif est minimisée.

Dans le cadre de la surveillance post-thérapeutique, il n'y a pas à notre connaissance de méthodologie statistique pour déterminer un calendrier optimal de visites et les calendriers utilisés en pratique reposent généralement sur des recommandations d'experts de sociétés scientifiques. Faciles à appliquer, ces calendriers ne détectent pas les récidives d'une manière optimale (Ataman et al. 2004). Beaucoup de récidives sont encore diagnostiquées après l'apparition des symptômes dans l'intervalle entre deux visites : Francken, Shaw, Neil et al. 2007 ; Francken, Shaw et J. F. Thompson 2008 ont constaté que dans le cadre des mélanomes, trois quart des récidives sont encore détectées par les patients ou leur entourage. Ces récidives symptomatiques sont diagnostiquées à des stades avancés, ce qui réduit considérablement le pronostic de guérison.

Plusieurs méthodologies ont été proposées dans la littérature pour individualiser le suivi du patient en fonction du risque de rechute. Wheeler et al. 1999 a proposé une méthode en deux étapes de planification des visites. Ils estiment dans un premier temps les risques de récidive annuels. Ils définissent ensuite un calendrier s'adaptant à ces estimations. Tsodikov et al. 1995 estime les délais avant récidive et les probabilités de diagnostic de faux positifs par une modélisation stochastique. Des méthodes bayésiennes ont aussi été proposées, notamment par Inoue et Parmigiani 2002 mais celles-ci sont

difficiles à mettre en œuvre. Kent et al. 1991 propose de planifier les cystoscopies pour détecter les récurrences du cancer de la vessie par une approche d'optimisation non linéaire. Cela aboutit à des intervalles inter-visites plus longs pour les patients à faible risque et des intervalles plus courts pour les patients avec un risque élevé. Enfin, Filleron et al. 2009 a proposé une stratégie en deux étapes : les facteurs pronostiques associés au délai de récidive sont identifiés et la fonction d'incidence cumulée est modélisée en fonction de ces facteurs de risque. La planification des visites se base alors sur les quantiles de cette fonction d'incidence cumulée.

Les méthodes existantes ne prennent pas en compte les différentes étapes de l'histoire naturelle de la maladie. Comme dans le cadre du dépistage, la surveillance post-thérapeutique est surtout utile quand les récidives sont détectées à des stades précoces pour lesquels le pronostic est meilleur (Peronne et al. 2004 ; Zola et al. 2007). Ainsi, un calendrier de surveillance post-thérapeutique optimal devrait maximiser la probabilité de détecter des récidives en phase préclinique pour les traiter avec succès. Un patient est dit au stade préclinique quand il n'a pas encore de symptôme de récidive : dans ce cas, la récidive ne peut être détectée que par des examens spécifiques. Pour la plupart des types de récidive, les traitements sont plus efficaces lorsque le patient est pris en charge à ce stade. Un calendrier optimal devrait aussi minimiser les visites inutiles. Enfin, un autre aspect spécifique de la phase post-thérapeutique est que le patient est à risque de plusieurs types d'événements (récidive locale, second cancer, métastases à distance...), qui présentent des caractéristiques différentes et doivent être modélisés différemment. La pertinence de la surveillance dépend également du type de récidive : par exemple, une détection précoce de l'apparition de métastases a peu d'importance en l'absence d'options curatives pour ce type de rechute.

Nous proposons dans l'article [11] un outil de prise de décision pour déterminer le calendrier optimal de visites : pour un nombre total de visites fixé et pour une durée totale de suivi fixée, notre calendrier permet de détecter les éventuelles récidives aussi précocement que possible dans la phase préclinique. La méthode distingue les différents types de récidives afin de donner la priorité aux types d'événements de meilleur pronostic. La méthode adapte la fonction d'utilité de Zelen 1993 proposée dans le cadre du dépistage des cancers. La section suivante présente le modèle de Zelen pour le dépistage puis notre adaptation de ce modèle au cas de plusieurs types de récidives est décrite. Nous avons appliqué la méthode au suivi des patients atteints de cancer du larynx. Enfin, quelques perspectives seront données.

Le modèle de Zelen pour le dépistage

L'approche de Zelen 1993 a pour objectif principal de maximiser la proportion de patients dont le cancer est détecté au stade asymptomatique lors des contrôles de dépistage et de minimiser la proportion de patients dont le cancer est diagnostiqué cliniquement dans l'intervalle entre deux visites planifiées. Zelen se base sur l'histoire naturelle du cancer, qui peut être décrite par trois états successifs, comme indiqué sur la figure 3.9 : à l'état initial, noté S_0, le patient n'est pas malade. Cet état inclut aussi les patients malades qui n'ont aucun symptôme et dont la maladie ne peut pas être détectée en l'état actuel des techniques médicales. S_p correspond au stade préclinique : le patient est atteint du cancer à un stade asymptomatique et celui-ci ne peut être détecté que par des examens spécifiques. Sans traitement, le patient va évoluer de cet état préclinique vers l'état clinique, noté S_c. A ce stade, le patient présente des symptômes évidents de la maladie qui peut alors être détectée par un simple examen médical. Le but du dépistage est donc d'effectuer des examens pour les individus de S_0 pour détecter dès que possible leur transition au stade S_p, avant qu'ils n'aient atteint le stade S_c.

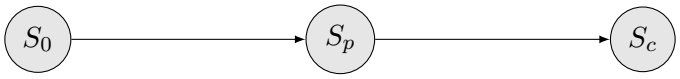


FIGURE 3.9 – Modèle d'histoire naturelle pour le dépistage.

Soit $[0,T_{max}]$ l'intervalle dans lequel sont programmées les $n+1$ visites de dépistage aux dates successives $0 = t_0 < t_1 < t_2 < \cdots < t_n = T_{max}$. Soit β la sensibilité de l'examen effectué pendant les séances de dépistage, c'est-à-dire la probabilité que l'examen détecte une tumeur présente.

Zelen définit $D_r(\beta)$ pour $r = 1,2,\ldots,n$ comme étant la probabilité qu'un cancer soit détecté en phase préclinique chez un patient lors de la séance de dépistage organisée à la date t_r, sachant la sensibilité β. Zelen définit ensuite $I_r(\beta)$ comme la probabilité, pour un individu participant à un programme de dépistage avec une sensibilité β et qui n'a pas été détecté au stade préclinique, de transiter de la phase préclinique à la phase clinique entre les visites t_{r-1} et t_r. L'objectif du dépistage étant de maximiser les probabilités $D_r(\beta)$ et de minimiser les probabilités $I_r(\beta)$, Zelen propose la fonction d'utilité suivante, qui est une combinaison linéaire de ces deux types de probabilités pour un choix judicieux des coefficients A_0, A et B :

$$U_{n+1}(\beta,T) = A_0 D_0(\beta) + A\sum_{r=1}^{n} D_r(\beta) - B\sum_{r=1}^{n} I_r(\beta). \tag{3.39}$$

L'interprétation de cette fonction dépend du choix des coefficients : par exemple, si ces coefficients représentent les probabilités de guérison selon le stade auquel est fait le diagnostic, alors U_{n+1} correspondra à la différence des taux de guérison entre les patients diagnostiqués aux visites planifiées et les patients qui ont transité vers le stade clinique.

La maximisation de cette fonction d'utilité permet de déterminer les n dates t_r $(r = 1,2,\ldots,n)$ optimales pour la planification du dépistage.

Application à la surveillance post-thérapeutique

Nous proposons dans cette section une adaptation de la méthode de Zelen qui nous permet de déterminer un calendrier optimal de surveillance post-thérapeutique en prenant en compte plusieurs types de récidives.

Notations

Durant leur suivi post-thérapeutique, les patients sont à risque de K types d'événements (récidive locale ou régionale, second cancer, métastases à distance...). Pour chaque type d'événement k $(k = 1,\ldots,K)$, un état préclinique S_{pk} (stade asymptomatique) précède un état clinique S_{ck} (stade symptomatique), comme le montre la figure 3.10. S_0 désigne toujours l'état sain. Dans ce modèle multi-états, nous désignons par $\omega_k(t)$ l'intensité de transition de S_0 vers S_{pk}, où t est le temps écoulé depuis l'entrée du patient en phase post-thérapeutique. $f_k(t)$ est la densité associée. L'intensité de transition vers un quelconque état préclinique vaut donc $\omega(t) = \sum_{k=1}^{K} \omega_k(t)$. L'intensité de transition de S_{pk} vers S_{ck} est notée $\lambda_k(z)$, où z est le temps écoulé depuis l'entrée du patient dans le stade préclinique associé à une récidive de type k. La densité et la fonction de survie associées sont notées respectivement $q_k(z)$ et $Q_k(z)$.

Les patients sont suivis pendant une durée fixée de T_{max} mois avec n visites potentielles planifiées à des temps fixes t_r $(r = 1,2,\ldots,n$ avec $0 < t_1 < \cdots < t_n = T_{max})$. Le $r$ème intervalle est l'intervalle $[t_{r-1},t_r[$. A chaque visite planifiée, des examens spécifiques

sont conduits dans le but de détecter les récidives en phase préclinique. Nous notons β_k les sensibilités des examens pour détecter une récidive de type k au stade préclinique. $D_{rk}(\beta_k)$ désigne la probabilité de détecter une récidive de type k au stade préclinique à la $r$ème visite et $I_{rk}(\beta_k)$ la probabilité qu'un patient en phase de surveillance transite de la phase préclinique pour l'événement de type k vers la phase clinique associée dans le $r$ème intervalle.

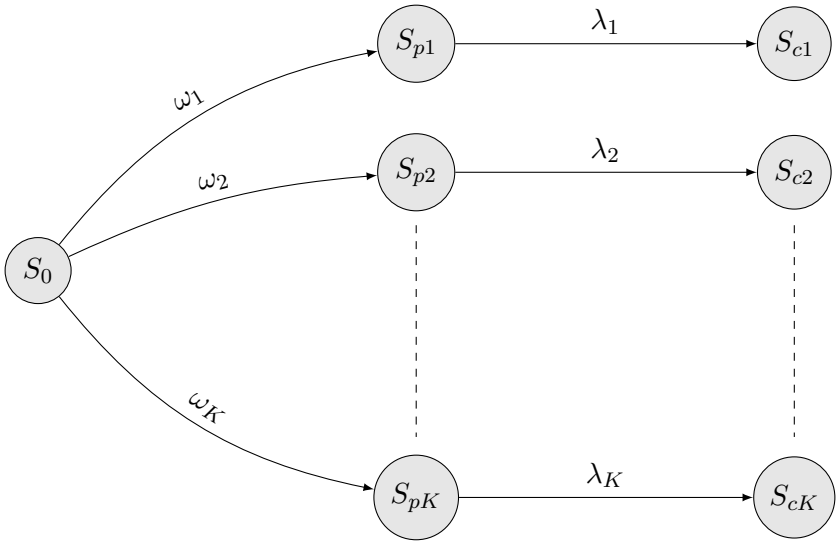


FIGURE 3.10 – Modèle d'histoire naturelle pour la surveillance post-thérapeutique.

Hypothèses simplificatrices

La figure 3.10 présente les $2K+1$ états du modèle ainsi que les intensités de transition associées. Nous faisons les hypothèses suivantes sur ces intensités de transition :

- les différents types d'événements sont considérés comme mutuellement exclusifs, si bien qu'un patient ne peut avoir qu'un seul type d'événement. En fait, dès qu'un type de récidive est détecté, le patient retourne en phase de traitement et quitte la phase de surveillance post-thérapeutique : ainsi, l'apparition d'une récidive d'un autre type peut être ignorée dans le modèle ;
- un patient dans un stade préclinique d'un événement de type donné n'est à risque de transiter que vers le stade clinique associé à ce type d'événement. Cette hypothèse paraît raisonnable étant donné que l'objectif est de détecter suffisamment tôt les récidives avant qu'elles ne transitent vers le stade clinique.

Notre modèle se base sur une hypothèse de chaînes de Markov homogènes (voir UHRY et al. 2010) dont les propriétés sont les suivantes :

- les intensités de transition sont invariantes avec le temps : $\omega_k(t) = \omega_k$, $\lambda_k(z) = \lambda_k$;
- les distributions des durées de séjour dans les différents états sont définies par $f_k(t) = \omega_k e^{-\omega t}$ avec $\omega = \sum_{k=1}^{K} \omega_k$ et $q_k(z) = \lambda_k e^{-\lambda_k z}$. Les durées de séjour dans les états successifs sont indépendantes ;
- les intensités de transition d'un état à un autre sont indépendantes des états passés et indépendantes du temps passé dans ces états.

La fonction d'utilité

Les probabilités $D_{rk}(\beta_k)$ et $I_{rk}(\beta_k)$ peuvent alors être calculées et il alors possible, à la manière de ZELEN 1993, de construire une fonction d'utilité combinaison linéaire de

ces probabilités qui maximise la probabilité de détection des récidives à un stade précoce et minimise la probabilité de transition vers un stade tardif. On pondère ces probabilités avec les coefficients de pondération A_k et B_k. Le coefficient A_k choisi (respectivement B_k) est la probabilité, pour un patient donné, d'être traité avec succès de sa récidive de type k détectée au stade préclinique (respectivement au stade clinique).

Pour un suivi prévoyant n visites sur une durée totale T_{max}, avec une sensibilité β_k de détecter une récidive de type k au stade préclinique, la fonction d'utilité est alors définie par :

$$U_n(\beta,t_1,\ldots,t_{n-1},T_{max}) = \sum_{r=1}^{n} A' D_r - \sum_{r=1}^{n} B' I_r, \tag{3.40}$$

où $\beta = (\beta_k)$, $A = (A_k)$, $B = (B_k)$, $I_r = (I_{rk}(\beta_k))$ et $D_r = (D_{rk}(\beta_k))$ sont des vecteurs à K composantes.

Cette fonction d'utilité peut être maximisée par l'algorithme du simplex proposé par Nelder et Mead 1964. Les valeurs t_r qui maximisent cette fonction sont les dates optimales auxquelles les visites doivent être programmées. En l'absence de perdus de vue, la méthode proposée permet aussi d'estimer le nombre de patients encore présents dans la phase de surveillance après un nombre de visites déjà effectué. On peut donc en déduire l'estimation du nombre total de visites à effectuer pour un planning de surveillance donné.

Illustration : le cancer du larynx

Pour le cancer du larynx, la stratégie de surveillance la plus communément appliquée en Europe est celle de la Société Européenne d'Oncologie Médicale (esmo) (voir Haas et al. 2001 ; Ritoe, Krabbe et al. 2004 ; Ritoe, De Vegt et al. 2007). La Fédération Nationale française des Centres de Lutte Contre le Cancer (fnclcc) propose sa propre variante. Enfin, le calendrier de surveillance le plus utilisé aux États-Unis est celui du National Comprehensive Cancer Network (nccn). Ces trois stratégies sont présentées dans le tableau 3.3 pour une durée de surveillance de 10 ans.

	Année de surveillance post-thérapeutique					
Société	1	2	3	4	5	6–10
esmo (31 visites)	1 visite par mois	1 visite tous les 2 mois	1 visite tous les 3 mois	1 visite tous les 6 mois	1 visite tous les 6 mois	1 visite tous les 12 mois
fnclcc (20 visites)	1 visite par mois pendant 6 mois puis tous les 3 mois	1 visite tous les 6 mois	1 visite tous les 6 mois	1 visite tous les 6 mois	1 visite tous les 12 mois	1 visite tous les 12 mois
nccn (18 visites)	1 visite tous les 3 mois	1 visite tous les 4 mois	1 visite tous les 6 mois	1 visite tous les 6 mois	1 visite tous les 6 mois	1 visite tous les 12 mois

Table 3.3 – Stratégies de surveillance post-thérapeutique pour le cancer du larynx proposées par trois sociétés scientifiques.

La méthodologie d'optimisation du calendrier de surveillance a été appliquée à une population de patients traités d'un cancer du larynx. Après leur traitement, les patients sont considérés à risque de trois types de récidives : récidives loco-régionales, métastases à distance et second cancer primaire. Les paramètres que nous avons utilisés pour le modèle de Markov multi-états et les coefficients de pondération A_k et B_k ont été trouvés dans la littérature (Ritoe, De Vegt et al. 2007) et sont donnés dans le tableau 3.4. Les sensibilités β_k des examens ont été fixées à 80 % pour les trois types de récidives en suivant les recommandations d'experts.

Les fonctions d'utilité ont été calculées pour les trois calendriers recommandés par les sociétés scientifiques et sont données dans le tableau 3.5. En maximisant la fonction

		Récidives loco-régionales	Métastases à distance	Second cancer primaire
Intensité $S_0 \to S_{pk}$ ($\times 10^{-3}$)	ω_k	3,250	1,667	9,500
Intensité $S_{pk} \to S_{ck}$	λ_k	0,350	0,317	0,149
Guérison après S_{pk}	A_k	0,737	0,000	0,923
Guérison après S_{ck}	B_k	0,803	0,000	0,440
Sensibilité	β_k	0,800	0,800	0,800

TABLE 3.4 – Paramètres pour le modèle de Markov multi-états du cancer du larynx.

d'utilité pour une durée de surveillance de 10 ans et en gardant le même nombre de visites fixé par chaque société, nous avons obtenu les calendriers optimaux de surveillance. A chaque fois, le calendrier optimal permettrait d'aboutir à une meilleure utilité et à un moindre nombre de visites. Enfin, nous proposons aussi un calendrier qui vise à réduire le nombre de visites par patient en conservant une utilité équivalente à celle du calendrier recommandé. Ainsi, les recommandations de l'ESMO pourraient être réarrangées en 25 visites (au lieu de 31) pour une utilité équivalente. Pour la FNCLCC, on passerait de 20 à 17 visites. On n'obtient pas de gain en nombre de visites planifiées par patient pour le NCCN mais en réorganisant les dates, on obtiendrait un nombre total de visites réduit.

		Nombre de visites par patient	Nombre total de visites pour 1000 patients	Proportion de guéris avec détection précoce	Proportion de guéris avec détection tardive	Utilité
ESMO	Recommandé	31	24108	0,40	0,18	0,22
	Optimal	31	20804	0,44	0,16	0,28
	Réduit	25	16836	0,41	0,18	0,23
FNCLCC	Recommandé	20	15113	0,33	0,23	0,11
	Optimal	20	13649	0,37	0,20	0,17
	Réduit	17	11675	0,34	0,22	0,12
NCCN	Recommandé	18	12685	0,35	0,22	0,13
	Optimal	18	12287	0,35	0,22	0,14
	Réduit	18	12287	0,35	0,22	0,14

TABLE 3.5 – Comparaison des caractéristiques des calendriers recommandés par trois sociétés scientifiques et des calendriers optimaux (même nombre de visites par patient) et réduits (utilité équivalente) associés.

Perspectives

Nous avons considéré ici un modèle de Markov homogène, ce qui est une hypothèse forte. Il est bien évident que le risque de récidive peut varier avec le temps. De plus, les intensités de transition d'un état à un autre dépendent en réalité du temps passé dans l'état de départ. Les intensités de transition peuvent aussi différer selon les caractéristiques des patients. A l'ère de la médecine personnalisée, il serait pertinent d'adapter le suivi aux patients ou au moins à des groupes homogènes de patients.

Nous avons également considéré les coefficients de pondération A_k et B_k comme fixés alors qu'il est fort probable que ces probabilités de guérison dépendent du temps écoulé depuis la fin du traitement. Ils peuvent bien sûr également dépendre des caractéristiques des patients.

Enfin, une extension possible de cette approche consiste à considérer un modèle dynamique, qui permet de considérer l'évolution de la santé du patient pendant sa phase de surveillance en l'autorisant à passer d'un groupe à un autre. Des techniques dont nous pourrions nous inspirer pour cela sont le modèle retardé de Markov proposé par ÖZEKICI et PLISKA 1991 ou les équations différentielles stochastiques fournies par VEESTRAETEN

2006.

3.3.7 Algorithme d'évaluation des stratégies de surveillance

Nous venons de voir que plusieurs stratégies peuvent être proposées pour la surveillance d'un cancer donné. L'évaluation des stratégies par un essai clinique, afin de déterminer la plus efficace, poserait cependant plusieurs difficultés tant éthiques que logistiques. Nous avons donc proposé dans la prépublication [23] un algorithme permettant d'effectuer cette évaluation au moyen d'outils de simulation numérique.

Les principes de modélisation dynamique des transitions (Siebert et al. 2012) sont appliqués pour la génération des histoires de la maladie des patients. La technique de génération utilisée est la simulation d'événements discrets (Fishman 2001 ; Karnon et al. 2012). Enfin la stratégie adoptée est orientée patient (H. T. O. Davies et R. Davies 1995).

3.4 Références

Aebi, S., T. Davidson, G. Gruber et M. Castiglione (2010). « Primary Breast Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up ». In : *Annals of Oncology* 21.5, p. 9–14.

Ambrogi, F., L. Trevisi, G. Martelli et P. Boracchi (2014). « Is breast cancer curable: a study of long-term crude cumulative incidence ». In : *Tumori* 100, p. 106–414.

Andersen, P. K., O. Borgan, R. D. Gill et N. Keiding (1993). *Statistical Models Based on Counting Processes*. New-York : Springer-Verlag.

Andersen, P. K. et R. D. Gill (1982). « Cox's Regression Model for Counting Processes: A Large Sample Study ». In : *The Annals of Statistics* 10.4, p. 1100–1120. URL : http://www.jstor.org/stable/2240714.

Arriagada, R., L. E. Rutqvist et A. Kramar (1992). « Competing risks determining event-free survival in early breast cancer ». In : *Br. J. Cancer* 66, p. 951–957. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1977968/.

Ataman, O. U., A. Barett, S. Davidson, D. De Haas-Kock, S. Dische, B. Dubray, I. M. Grillo, A. Kramar, C. Haie-Meder, G. Heeren, K. Hideghety, J. LeVay, J. Maher, M. Marcenaro, R.-P. Muller, C. A. Reguerio, M. I. Saunders, I. Turesson, P. Van Houtte et V. Vitale (2004). « Audit of effectiveness of routine follow-up clinics after radiotherapy for cancer: a report of the REACT Working Group of ESTRO ». In : *Radiotherapy and oncology* 73.2, p. 237–249. URL : http://www.sciencedirect.com/science/article/pii/S0167814004002063.

Bagnoli, M. et T. Bergstrom (2005). « Log-Concave Probability and Its Applications ». In : *Economic Theory* 26.2, p. 445–469. URL : http://www.jstor.org/stable/25055959.

Balmaña, J., A. Castells et A. Cervantes (2010). « Familial Colorectal Cancer Risks: ESMO Clinical Practice Guidelines ». In : *Clinical Oncology* 21.5, p. 78–81.

Beitler, A. L., K. S. Virgo, F. E. Johnson, J. F. Gibbs et W. G. Kraybill (2000). « Current follow-up strategies after potentially curative resection of extremity sarcomas: results of a survey of the members of the Society of Surgical Oncology ». In : *Cancer* 88.4, p. 777–785.

Betensky, R. A. et D. A. Schoenfeld (2001). « Nonparametric Estimation in a Cure Model with Random Cure Times ». In : *Biometrics* 57.1, p. 282–286. URL : http://www.jstor.org/stable/2676872.

Biggs, M. L. et S. M. Schwartz (2007). *Cancer of the Testis*. National Cancer Institute. SEER Program, NIH Pub. No. 07-6215. URL : http://seer.cancer.gov/archive/publications/survival/seer_survival_mono_lowres.pdf.

Boag, J.W. (1949). « Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 11.1, p. 15–53. URL : http://www.jstor.org/stable/2983694.

Byar, D. P. (1980). « The Veterans Administration Study of Chemoprophylaxis for Recurrent Stage I Bladder Tumours: Comparisons of Placebo, Pyridoxine and Topical Thiotepa ». In : *Pavone-Macaluso M., Smith P. H., Edsmyr F. (eds), Bladder Tumors and other Topics in Urological Oncology. Ettore Majorana International Science Series (Life Sciences), Springer, Boston, MA* 1, p. 363–370. DOI : https://doi.org/10.1007/978-1-4613-3030-1_74.

Cai, C., Y. Zou, Y. Peng et J. Zhang (2012). « smcure : An R-package for estimating semiparametric mixture cure models ». In : *Computer Methods and Programs in Biomedicine* 108.3, p. 1255–1260. DOI : 10.1016/j.cmpb.2012.08.013.

Cardoso, F., E. Senkus-Konefka, L. Fallowfield, A. Costa et M. Castiglione (2010). « Locally Recurrent and Metastatic Breast Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up ». In : *Annals of Oncology* 21.5, p. 15–19.

Casali, P. G., J. Y. Blay et ESMO Guidelines Working Group (2010). « Soft Tissue Sarcomas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up ». In : *Annals of Oncology* 21.5, p. 198–203.

Clayton, D. et J. Cuzick (1985). « Multivariate Generalizations of the Proportional Hazards Model ». In : *Journal of the Royal Statistical Society. Series A (General)* 148.2, p. 82–117. URL : http://www.jstor.org/stable/2981943.

Coindre, J. M. (2006). « Grading of soft tissue sarcomas, review and update ». In : *Archives of Pathology and Loboratory Medicine* 130, p. 1448–1453.

Cook, R. J. et J. F. Lawless (1997). « Marginal Analysis of Recurrent Events and a Terminating Event ». In : *Statistics in Medicine* 16.8, p. 911–924. URL : http://dx.doi.org/10.1002/(SICI)1097-0258(19970430)16:8%3C911::AID-SIM544%3E3.0.CO;2-I.

Corbière, F. et P. Joly (2007). « A SAS macro for parametric and semiparametric mixture cure models ». In : *Computer Methods and Programs in Biomedicine* 85.2, p. 173–180. URL : http://www.sciencedirect.com/science/article/pii/S0169260706002513.

Cox, D. R. (1972). « Regression Models and Life-Tables ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2, p. 187–220. URL : http://www.jstor.org/stable/2985181.

Culine, S., P. Kerbrat, A. Kramar, C. Théodore, C. Chevreau, L. Geoffrois, N. B. Bui, J. Pény, A. Caty, R. Delva, P. Biron, K. Fizazi, J. Bouzy et J.-P. Droz (2007). « Refining the optimal chemotherapy regimen for good-risk metastatic nonseminomatous germ-cell tumors: a randomized trial of the Genito-Urinary Group of the French Federation of Cancer Centers (GETUG T93BP) ». In : *Annals of Oncology* 18.5, p. 917–924. URL : http://annonc.oxfordjournals.org/content/18/5/917.

Culine, S., A. Kramar, C. Théodore, L. Geoffrois, C. Chevreau, P. Biron, B. B. Nguyen, J.-F. Héron, P. Kerbrat, A. Caty, R. Delva, P. Fargeot, K. Fizazi, J. Bouzy et J.-P. Droz (2008). « Randomized Trial Comparing Bleomycin/Etoposide/Cisplatin With Alternating Cisplatin/Cyclophosphamide/Doxorubicin and Vin-

blastine/Bleomycin Regimens of Chemotherapy for Patients With Intermediate- and Poor-Risk Metastatic Nonseminomatous Germ Cell Tumors: Genito-Urinary Group of the French Federation of Cancer Centers Trial T93MP ». In : *Journal of Clinical Oncology* 26.3, p. 421–427. URL : http://jco.ascopubs.org/content/26/3/421.

DAVIES, H. T. O. et R. DAVIES (1995). « Simulating health systems: modelling problems and software solutions ». In : *European Journal of Operational Research* 87.1, p. 35–44. URL : http://www.sciencedirect.com/science/article/pii/0377221794001486.

DESCH, C. E., A. B. BENSON, M. R. SOMERFIELD, P. J. FLYNN, C. KRAUSE, C. L. LOPRINZI, B. D. MINSKY, D. G. PFISTER, K. S. VIRGO et N. J. PETRELLI (2005). « Colorectal Cancer Surveillance: 2005 Update of an American Society of Clinical Oncology Practice Guideline ». In : *Journal of Clinical Oncology* 23.33, p. 8512–8519. URL : http://jco.ascopubs.org/content/23/33/8512.

DICKMAN, P. W., A. SLOGGETT, M. HILLS et T. HAKULINEN (2004). « Regression models for relative survival ». In : *Statistics in Medicine* 23, p. 51–64.

DRAISMA, G. et J. VAN ROSMALEN (2013). « A note on the catch-up time method for estimating lead or sojourn time in prostate cancer screening ». In : *Statistics in Medicine* Published online.

EDERER, F., L. M. AXTELL et S. J. CUTLER (1961). « The relative survival rate: a statistical methodolgy ». In : *National Cancer Institute Monograph* 6, p. 101–121.

FERLAY, J., I. SOERJOMATARAM, M. ERVIK, D. FORMAN, F. BRAY, R. DIKSHIT, S. ELSER et C. MATHERS (2013). *Cancer Incidence and Mortality Worldwide in 2012*. International Agency for Research on Cancer, World Health organization. Globocan 2012. URL : http://globocan.iarc.fr/factsheet.asp.%20Consult%7B%5C'e%7D%20le%2003%20avril%202012.

FILLERON, T., A. BARETT, O. U. ATAMAN et A. KRAMAR (2009). « Planning post-therapeutic oncologic surveillance visits based on individual risks ». In : *Medical Decision Making* 29.5, p. 570–579.

FISHMAN, G. S. (2001). *Discrete-Event Simulation: Modeling, Programming, and Analysis*. Springer. ISBN : 9780387951607.

FRANCKEN, A. B., H. M. SHAW, A. A. NEIL, S.-J. SOONG, H. J. HOEKSTRA et J. F. THOMPSON (2007). « Detection of first relapse in cutaneous melanoma patients: Implication for the formulation of evidence-based follow-up guidelines ». In : *Annals of Surgical Oncology* 14.6, p. 1924–1933.

FRANCKEN, A. B., H. M. SHAW et J. F. THOMPSON (2008). « Detection of second primary cutaneous melanoma ». In : *European Journal of Surgical Oncology* 34, p. 587–592.

GAMEL, J.W. et R.L. VOGEL (1997). « Comparison of parametric and non-parametric survival methods using simulated clinical data ». In : *Statistics in medicine* 16.14, p. 1629–1643.

GENT, M., D. BEAUMONT, J. BLANCHARD, M. G. BOUSSER, J. COFFMAN, J. D. EASTON, J. R. HAMPTON, L. A. HARKER, L. JANZON, J. J. E. KUSMIEREK, E. PANAK, R. ROBERTS, J. S. SHANNON, J. SICURELLA, G. TOGNONI, E. J. TOPOL, M. VERSTRAETE et C. WARLOW (1996). « A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE) ». In : *The Lancet* 348, p. 1329–1339.

GERRAND, C. H., L. J. BILLINGHAM, P. J. WOLL et R. J. GRIMER (2007). « Follow up after Primary Treatment of Soft Tissue Sarcoma: A Survey of Current Practice in the United Kingdom ». In : *Sarcoma*. DOI : 10.1155/2007/34128.

Ghosh, D. et D. Y. Lin (2000). « Nonparametric Analysis of Recurrent Events and Death ». In : *Biometrics* 56.2, p. 554–562. URL : http://dx.doi.org/10.1111/j.0006-341X.2000.00554.x.

Gompertz, B. (1825). « On the nature of the function expressive of the law of human mortality, and on a new mode on determining the value of life contingencies ». In : *Philosophical Transactions of the Royal Society of London* 115, p. 513–80.

Gooley, T.A., W. Leisenring, J. Crowley et B. E. Storer (1999). « Estimation of failure probabilities in the presence of competing risks: new representations of old estimators ». In : *Statistics in Medicine* 18.6, p. 695–706. URL : http://dx.doi.org/10.1002/(SICI)1097-0258(19990330)18:6%3C695::AID-SIM60%3E3.0.CO;2-O.

Gray, R. J. (2011). « cmprsk: Subdistribution analysis of competing risks ». In : URL : http://CRAN.R-project.org/package=cmprsk/.

Haas, I., U. Hauser et U. Ganzer (2001). « The dilemma of follow-up in head and neck cancer patients ». In : *European Archive of Otorhinolaryngology* 258, p. 177–183.

Halmos, P. R. et L. J. Savage (1949). « Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics ». In : *Ann. Math. Statist.* 20.2, p. 225–241. URL : https://doi.org/10.1214/aoms/1177730032.

Hofmann, U., M. Szedlak, W. Rittgen, E. G. Jung et D. Schadendorf (2002). « Primary staging and follow-up in melanoma patients - monocenter evaluation of methods, costs and patients survival ». In : *British Journal of Cancer* 87, p. 151–157.

Hogendoorn, P. C. W., N. Athanasou, S. Bielack, E. De Alava, A. P. D Tos, S. Ferrari, H. Gelderblom, R. Grimer, K. S. Hall, B. Hassan, H. Jurgens, M. Paulussen, L. Rozeman, A. H .M. Taminiau, J. Whelan et D. Vanel (2010). « Bone sarcomas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up ». In : *Annals of Oncology* 21.suppl 5, p. v204–v213. URL : http://annonc.oxfordjournals.org/content/21/suppl_5/v204.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer-Verlag New York. ISBN : 978-0-387-98873-3.

Hudgens, M., C. Li et J. Fine (2011). « Parametric Estimation of the Cumulative Incidence Function for Interval Censored Competing Risks Data ». In : *The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series* 23.

Inoue, L. Y. T. et G. Parmigiani (2002). « Designing Follow-Up Times ». In : *Journal of the American Statistical Association* 97.459, p. 847–858. URL : http://dx.doi.org/10.1198/016214502388618645.

Jeong, J.-H. et J. P. Fine (2006). « Direct parametric inference for the cumulative incidence function ». In : *Applied Statistics* 55.2, p. 187–200.

— (2007). « Parametric regression on cumulative incidence function ». In : *Biostatistics* 8, p. 184–196.

Kalbfleisch, J. D. et R. L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. John Wiley et Sons, Inc. ISBN : 9781118032985.

Kaplan, E. L. et P. Meier (1958). « Nonparametric Estimation from Incomplete Observations ». In : *Journal of the American Statistical Association* 53.282, p. 457–481. URL : http://www.jstor.org/stable/2281868.

Karnon, J., J. Stahl, A. Brennan, J. J Caro, J. Mar et J. Möller (2012). « Modeling Using Discrete Event Simulation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force–4 ». In : *Medical Decision Making* 32.5, p. 701–711. URL : http://mdm.sagepub.com/content/32/5/701.

Kent, D. L., R. A. Nease, H. C. Sox, L. D. Shortliffe et R. Shachter (1991). « Evaluation of Nonlinear Optimization for Scheduling of follow-up cystoscopies to Detect Recurrent Bladder Cancer ». In : *Medical Decision Making* 11.4, p. 240–248. doi : https://doi.org/10.1177/0272989X9101100402.

Khatcheressian, J. L., A. C. Wolff, T. J. Smith, E. Grunfeld, H. B. Muss, V. G. Vogel, F. Halberg, M. R. Somerfield et N. E. Davidson (2006). « American Society of Clinical Oncology 2006 Update of the Breast Cancer Follow-Up and Management Guidelines in the Adjuvant Setting ». In : *Journal of Clinical Oncology* 24.31, p. 5091–5097. url : http://jco.ascopubs.org/content/24/31/5091.

Lai, X et K. K. Yau (2009). « Multilevel mixture cure model with random effects ». In : *Biometrical Journal* 51.3, p. 456–66.

Lambert, P. C. (2007). « Modeling of the cure fraction in survival studies ». In : *Stata Journal* 7.3, p. 351. url : http://www.pauldickman.com/workshop/curemodels_statajournal.pdf.

Lau, B., S. R. Cole et S. J. Gange (2011). « Parametric mixture models to evaluate and summarize hazard ratios in the presence of competing risks with time-dependent hazards and delayed entry ». In : *Statistics in Medicine* 30.6, p. 654–665.

Lawless, J. F. et C. Nadeau (1995). « Some Simple Robust Methods for the Analysis of Recurrent Events ». In : *Technometrics* 37.2, p. 158–168. url : http://www.jstor.org/stable/1269617.

Lee, S. J. et M. Zelen (1998). « Scheduling periodic examinations for the early detection of disease: Application to breast cancer ». In : *Journal of the American Statistical Association* 93.444, p. 1271–1281.

Lee, S. Y., S. H. Jeong, J. Kim, S. H. Jung, K. B. Song et C. M. Nam (2007). « Scheduling mammography screening for early detection of breast cancer in Korean women ». In : *Journal of Medical Screening* 14, p. 205–209.

Menjoge, S. S. (2003). « On estimation of frequency data with censored observations ». In : *Pharmaceutical Statistics* 2.3, p. 191–197. url : http://dx.doi.org/10.1002/pst.37.

Mould, R. F., B. Asselain et Y. De Rycke (2004). « Methodology to predict maximum follow-up period for breast cancer patients without significantly reducing the chance of detecting local recurrence ». In : *Physics in Medicine and Biology* 49, p. 1079–1083.

Nelder, J. A. et R. Mead (1964). « A simplex algorithm for function minimization ». In : *Computer Journal* 7, p. 148–154.

Özekici, S. et S. R. Pliska (1991). « Optimal scheduling of inspections: A delayed Markov model with false positives and negatives ». In : *Operations Research* 39.2, p. 261–273.

Peng, Y., K. B. G. Dear et J. W. Denham (1998). « A generalized F mixture model for cure rate estimation ». In : *Statistics in Medicine* 17, p. 813–830.

Peng, Y. et J. R. G. Taylor (2010). « Mixture cure model with random effects for the analysis of multi-center toncil cancer study ». In : *Statistics in Medicine* 30, p. 211–23.

Peng, Y. et J. Zhang (2008). « Estimation method for the semiparametric mixture cure gamma frailty model ». In : *Statistics in Medicine* 27, p. 5177–94.

Peronne, M. A., A. Musolino, M. Michiara, B. Di Blasio, V. Franciosi, G. Cocconi, R. Camisa, R. Todeschini et S. Cascinu (2004). « Early detection of recurrences in the follow-up of primary breast cancer in an asymptomatic or symptomatic phase. » In : *Tumori* 90.3, p. 276–279.

Prentice, R. L., B. J. Williams et A. V. Peterson (1981). « On the regression analysis of multivariate failure time data ». In : *Biometrika* 68.2, p. 373–379. URL : http://dx.doi.org/10.1093/biomet/68.2.373.

Ritoe, S. C., F. De Vegt, I. M. Scheike, P. F. M. Krabbe, J. H. A. M. Kaanders, F. J. A. Van Den Hoogen, A. L. M. Verbeek et H. A. M. Marres (2007). « Effect of routine follow-up after treatment for laryngeal cancer on life expectancy and mortality. Results of a Markov model analysis ». In : *American Cancer Society* 109.2, p. 239–247.

Ritoe, S. C., P. F. M. Krabbe, J. H. A. M. Kaanders, F. J. A. Van Den Hoogen, A. L. M. Verbeek et H. A. M. Marres (2004). « Value of routine follow-up for patients cured of laryngeal carcinoma ». In : *Cancer* 101, p. 1382–1389.

Shen, Y. et G. Parmigiani (2003). *Optimization of breast cancer screening modalities*. URL : http://biostats.bepress.com/jhubiostat/paper18.

Shen, Y. et M. Zelen (2001). « Screening Sensitivity and sojourn time from breast cancer early detection clinical trials: mammograms and physical examinations. » In : *Journal of Clinical Oncology* 19, p. 3490–3499.

— (2005). « Robust modeling in screening studies: Estimation of sensitivity of preclinical sojourn time distribution ». In : *Biostatistics* 6.4, p. 604–614.

Siebert, U., O. Alagoz, A. M. Bayoumi, B. Jahn, D. K. Owens, D. J. Cohen et K. M. Kuntz (2012). « State-Transition Modeling A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force–3 ». In : *Medical Decision Making* 32.5, p. 690–700. URL : http://mdm.sagepub.com/content/32/5/690.

Sorensen, M., M. Pijls-Johannesma et E. Felip (2010). « Small-Cell Lung Cancer: ESMO Clinical Practices Guidelines for Diagnosis, Treatment and Follow-Up ». In : *Clinical Oncology* 21.5, p. 120–125.

Sposto, R. (2002). « Cure model analysis in cancer: an application to data from the Children's Cancer Group ». In : *Statistics in Medicine* 21.2, p. 293–312. URL : http://onlinelibrary.wiley.com/doi/10.1002/sim.987/abstract.

Steward, B. W. et C. P. Wild (2014). *World Cancer Report 2014*. IARC Nonserial Publication. URL : http://apps.who.int/bookorders/anglais/detart1.jsp?codlan=1&codcol=76&codcch=31.

Stewart-Merrill, S. B., S. A. Boorjian, R. H. Thompson, S. P. Psutka, J. C. Cheville, P. Thapa, E. J. Bergstrahl, M. K. Tollefson et I. Frank (2015). « Evaluation of current surveillance guidelines following radical cystectomy and proposal of a novel risk-based approach ». In : *Urologic Oncology: Seminars and Original Investigations* 33.8, 339.e1–339.e8. URL : http://dx.doi.org/10.1016/j.urolonc.2015.04.017.

Stewart-Merrill, S. B., R. H. Thompson, S. A. Boorjian, S. P. Psutka, C. M. Lohse, J. C. Cheville, B. C. Leibovich et I. Frank (2015). « Oncologic Surveillance After Surgical Resection for Renal Cell Carcinoma: A Novel Risk-Based Approach ». In : *Journal of Clinical Oncology*, JCO.2015.61.8009. URL : http://jco.ascopubs.org/content/early/2015/09/03/JCO.2015.61.8009.

Tai, Patricia, Edward Yu, Gábor Cserni, Georges Vlastos, Melanie Royce, Ian Kunkler et Vincent Vinh-Hung (2005). « Minimum follow-up time required for the estimation of statistical cure of cancer patients: verification using data from 42 cancer sites in the SEER database ». In : *BMC Cancer* 5.1, p. 48. URL : https://doi.org/10.1186/1471-2407-5-48.

Tsiatis, A (1975). « A nonidentifiability aspect of the problem of competing risks ». In : *Proceedings of the National Academy of Sciences* 72.1, p. 20–22. URL : http://www.pnas.org/content/72/1/20.abstract.

Tsodikov, A. D., B. Asselain, A. Fourque, T. Hoang et A. Y. Yakovlev (1995). « Discrete strategies of cancer post-treatment surveillance. Estimation and optimization problems ». In : *Biometrics* 51.2, 437–447. URL : http://www.imise.uni-leipzig.de/Archiv/1995/tsodikov-a-1995-437-a.pdf.

Uhry, Z., G. Hédelin, M. Colonna, B. Asselain, P. Arveux, A. Rogel, C. Exbrayat, I. Courtial, P. Soler-Michel, F. Molinié, D. Eilstein et S. W. Duffy (2010). « Multi-state Markov models in cancer screening evaluation: a brief review and case study ». In : *Statistical Methods in Medical Research* 19.5, p. 463–486.

Veestraeten, D. (2006). « An alternative approach to modelling relapse in cancer with an application to adenocarcinoma of the prostate ». In : *Mathematical Biosciences* 199.1, p. 38–54. URL : http://hinari-gw.who.int/whalecomwww.sciencedirect.com/whalecom0/science/article/pii/S0025556405001896.

Weaver, K. E., N. M. Aziz, N. K. Arora, L. P. Forsythe, A. S. Hamilton, I. Oakley-Girvan, G. Keel, K. M. Bellizzi et J. H. Rowland (2014). « Follow-Up Care Experiences and Perceived Quality of Care Among Long-Term Survivors of Breast, Prostate, Colorectal, and Gynecologic Cancers ». In : *Journal of Oncology Practice* 10.4, e231–e239. URL : http://jop.ascopubs.org/content/10/4/e231.

Wei, L. J., D. Y. Lin et L. Weissfeld (1989). « Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions ». In : *Journal of the American Statistical Association* 84.408, p. 1065–1073. URL : http://www.jstor.org/stable/2290084.

Wheeler, T., S. Stenning, S. Negus, S. Picken et S. Metcalfe (1999). « Evidence to support a change in follow-up policy for patients with breast cancer: time to first relapse and hazard rate analysis ». In : *Clinical Oncology* 11, p. 169–173.

who (2015). who / *Cancer*. URL : http://who.int/mediacentre/factsheets/fs297/en/.

Witten, M. et W. Satzer (1992). « Gompertz survival model parameters: Estimation and sensitivity ». In : *Applied Mathematics Letters* 5.1, p. 7–12.

Zelen, M. (1993). « Optimal scheduling of examinations for the early detection of disease ». In : *Biometrika* 80.2, p. 279–293.

Zhang, J. J. et M. Wang (2009). « An accelerated failure time mixture model with masked event ». In : *Biometrical Journal* 51.6, p. 932–45.

Zola, P., L. Fuso, S. Mazzola, E. Piovano, S. Perrotto, A. Gadducci, L. Galletto, F. Landoni, T. Maggino, F. Raspagliesi, E. Sartori et G. Scambia (2007). « Could follow-up different modalities play a role in asymptomatic cervical cancer relapses diagnosis? An italian multicenter retrospective analysis ». In : *Gynecologic Oncology* 107, p. 150–154.

4 — Sélection de variables

Les problèmes de sélection de variables connaissent un intérêt croissant avec le traitement de bases de données contenant de plus en plus de variables. Durant les vingt dernières années, de nombreuses méthodes de sélection de variables ont été proposées afin de traiter ces problèmes en grande dimension, en particulier quand le nombre p de variables dépasse le nombre n d'individus. Pour éviter une mauvaise estimation en raison de problèmes de colinéarité et améliorer l'interprétation, la communauté scientifique a développé des outils afin de sélectionner les variables les plus pertinentes. La littérature abonde dans le cas de la régression linéaire. Une méthode classique bien connue est l'algorithme pas-à-pas fondé sur le critère d'information d'Akaike (AIC). Récemment, des méthodes basées sur l'optimisation, telles que la pénalisation $\mathcal{L}_1$, ont été proposées. Des algorithmes basés sur les arbres de régression offrent aussi une alternative intéressante pour traiter non paramétriquement un grand nombre de covariables.

Dans le cadre de la thèse de Marie Walschaerts, sous la direction de Philippe Besse et de Patrick Thonneau, nous nous sommes intéressés dans la prépublication [19] au cas particulier de la sélection de modèles dans le cas où la variable réponse est censurée à droite. Nous avons recensé et comparé les nouvelles méthodes de sélection stable de variables basées sur le bootstrap pour deux méthodologies différentes communément utilisées dans les analyses de survie : le modèle à risques proportionnels de Cox et les arbres de survie. Ce sera l'objet de la section 4.1. Dans la section 4.2, nous proposons une signature en oncologie pour le cancer du poumon, basée sur cinq gènes (article [8]). Enfin, dans la section 4.3, nous considérons le problème de la sélection de modèles de survie en grande dimension dans le contexte des risques concurrents, qui était le sujet du stage de M2 de Soufiane Ajana (article [14]).

4.1 Comparaison de méthodes de sélection de modèles de survie

4.1.1 Motivation

Il est de plus en plus fréquent de disposer dans les analyses de survie de données en grande dimension. On cherche alors à sélectionner les variables les plus pertinentes afin d'obtenir les meilleurs modèles de prédiction possibles, en visant l'interprétabilité et la stabilité de ces modèles.

Le modèle de Cox (voir 1.5) est l'outil statistique le plus couramment utilisé pour modéliser la relation entre une durée de vie censurée et des facteurs de risque, en particulier dans le domaine médical. Ce modèle a l'avantage de ne pas nécessiter d'hypothèses sur la distribution des temps de survie. Il est facile d'interprétation pour les cliniciens et permet de fournir des estimations de l'effet des covariables sur la durée de survie après ajustement sur les autres covariables. Dans le cas d'un grand nombre de covariables, plusieurs approches statistiques sont à notre disposition pour réaliser la sélection du meilleur modèle dans le cadre du modèle de Cox, comme la méthode pas-à-pas basée sur l'AIC ou les méthodes basées sur la pénalisation $\mathcal{L}_1$ (méthode Lasso). Mais les modèles obtenus s'avèrent très souvent instables. L'instabilité des méthodes de sélection se rencontre également dans les modèles de régression linéaire (voir HARRELL, K. L. LEE et al. 1984 ; MEINSHAUSEN et BUEHLMANN 2010). Afin de pallier ces problèmes, certains auteurs ont proposé de combiner des techniques de ré-échantillonnage à ces méthodes de sélection de variables. Afin d'obtenir un modèle de sélection stable, BACH 2009 a proposé une version bootstrappée du Lasso, appelée Bolasso. D'autre part, MEINSHAUSEN et BUEHLMANN 2010 ont proposé une généralisation de la procédure Lasso, appelée Lasso randomisé. Ces méthodes Lasso bootstrappées ont uniquement été considérées dans le cadre de la régression linéaire (à la date de [19]) et nous proposons de les étendre au cas du modèle de Cox pour stabiliser la sélection des covariables.

D'autre part, des méthodes à base d'arbres de régression adaptées aux données de survie ont connu un essor important durant ces dernières décennies. Bien que moins connues que le modèle de Cox, elles offrent une bonne alternative pour identifier les variables qui jouent un rôle sur la survie et pour prédire le risque individuel d'un individu au vu de ses covariables. En plus d'être faciles à interpréter dans un large cadre d'applications, les arbres de survie peuvent intégrer des effets non linéaires et aussi prendre en compte les interactions entre les covariables sans avoir à les spécifier. Mais ces méthodes ont tendance à sur-ajuster les données et souffrent aussi d'instabilité (voir BREIMAN 1996 ; DANNEGGER 2000 ; RUEY-HSIA 2001 ; GEY et POGGI 2006). Une méthode de stabilisation bien connue pour améliorer la performance de prédiction d'un arbre unique consiste à agréger une famille d'arbres construits sur des échantillons bootstrap avec une sélection aléatoire des covariables à chaque nœud. Cette procédure, proposée par BREIMAN 1984 et appelée « forêts aléatoires », a été adaptée au domaine de la survie par ISHWARAN, KOGALUR et al. 2008. Les « forêts aléatoires de survie » sont considérées comme la meilleure modélisation en termes de performance prédictive mais sont beaucoup moins simples d'interprétation qu'un arbre unique pour les cliniciens. Par ailleurs, DANNEGGER 2000 a proposé une procédure de stabilisation par bootstrap au niveau de chaque nœud pour les arbres de survie. Cette méthodologie a l'avantage de conserver l'interprétation intuitive d'un arbre unique.

Dans ce travail, nous étendons les méthodes Bolasso et Bolasso randomisé au modèle de Cox, décrivons et comparons ces différentes méthodes de sélection stable de variables basées sur le modèle de Cox et sur les arbres de survie sur deux jeux de données réels : un jeu de données classique sur le cancer du sein et un jeu de données original sur l'infertilité masculine. Nous avons basé notre comparaison sur des critères statistiques tels que la qualité de la prédiction, mais aussi sur la capacité de la méthode à fournir un modèle final dont l'interprétation clinique est simple et pertinente.

Dans la section 4.1.2, une présentation succincte des méthodes de sélection de variables basées sur le modèle de Cox est donnée, suivie, dans la section 4.1.3, par la description des méthodes de sélection basées sur les arbres de survie. La section 4.1.4 compare les différentes approches sur les deux jeux de données, fournissant l'erreur de

prédiction et les variables retenues dans le modèle final pour chaque approche. Des conclusions et perspectives sont présentées en section 4.1.5. Les notations seront celles utilisées dans les précédents chapitres.

4.1.2 Méthodes de sélection de variables basées sur le modèle de Cox

Sélection pas-à-pas avec ré-échantillonnage (BSS)

Dans le cas d'un grand nombre de covariables, la sélection des prédicteurs dans un modèle de Cox est généralement effectuée par un algorithme pas-à-pas basé sur l'AIC (noté *stepwise*). Cependant, cette sélection de variables peut se révéler instable. En scindant l'échantillon, HARRELL, K. L. LEE et al. 1984 a montré que les prédicteurs sélectionnés n'étaient pas toujours les mêmes. Ce résultat a été confirmé par CHEN et GEORGE 1985 qui ont appliqué la procédure de pas-à-pas sur 100 échantillons bootstrappés. Seuls 2 % des modèles obtenus coïncident avec le modèle basé sur l'échantillon initial. Pour remédier à ce problème, SAUERBREI et SCHUMACHER 1992 ont mis au point une procédure de sélection qui combine la méthode du bootstrap et la sélection *stepwise*, que nous nommerons BSS (pour *Bootstrap Stepwise Selection*). Ils ont examiné les fréquences d'inclusion des variables sélectionnées dans les modèles ajustés sur les échantillons bootstrap et suggèrent de garder dans le modèle final les variables dont la fréquence d'inclusion est supérieure à un seuil κ choisi arbitrairement. Ils ont appliqué leur méthode sur un jeu de données de tumeurs cérébrales contenant 447 patients et 12 covariables. Ils ont montré que le choix de $\kappa = 0{,}6$ aboutissait au même modèle final quel que soit le nombre d'échantillons bootstrap.

Sélection Lasso bootstrappé et Lasso bootstrappé randomisé

Sélection Lasso

Une alternative à la procédure de sélection *stepwise* est la méthode de pénalisation $\mathcal{L}_1$ appelée aussi Lasso (*Least Absolute Shrinkage and Selection Operator*). Adaptée par TIBSHIRANI 1997 au modèle de Cox, cette méthode permet d'estimer les paramètres β via la maximisation de la log-vraisemblance partielle (1.5) sous la contrainte $\sum_{j=1}^{p} |\beta_j| \leq \lambda$ où λ est un paramètre de régularisation. La contrainte Lasso sélectionne les variables en enlevant celles qui ont les plus petits coefficients estimés. Cela conduit à des coefficients exactement égaux à zéro et permet d'obtenir un modèle parcimonieux et interprétable. Par contre, le choix du paramètre λ n'est pas anodin et MEINSHAUSEN et BUEHLMANN 2010 ont montré dans le cadre de la régression linéaire que le modèle obtenu était sensible à ce choix. Ils ont étudié la stabilité de la sélection de gènes sur un jeu de données en introduisant du bruit dans les covariables : les 4088 covariables sauf 6 d'entre elles ont été permutées. Quand la valeur de λ augmente, le modèle final retient les 6 gènes non permutés mais aussi les variables inadéquates. MEINSHAUSEN et BUEHLMANN 2010 montrent par ailleurs que la validation croisée pour le choix de λ n'est pas une bonne alternative pour les problèmes en grande dimension : dans leur exemple, 14 variables inadéquates sont sélectionnées. En outre, MEINSHAUSEN et BUEHLMANN 2006 ont montré que quand le nombre de covariables tend vers l'infini, la probabilité de sélectionner une covariable sans effet tend vers 1.

Sélection Lasso bootstrappé (BLS)

Pour obtenir une sélection stable du modèle, BACH 2009 a proposé dans le cadre du modèle linéaire une version bootstrappée du Lasso, appelée Bolasso. Le modèle final retenu est composé uniquement des variables qui ont été sélectionnées dans tous les échantillons bootstrap. Autrement dit, en employant la terminologie du modèle BSS, cela correspond à une valeur seuil κ égale à 1. Cependant, MEINSHAUSEN et BUEHLMANN 2010 ont montré sur la base de leur échantillon permuté que même en utilisant un sous-

échantillonnage aléatoire (une procédure proche du bootstrap, voir Freedman 1977), l'algorithme Lasso pouvait sélectionner des covariables non pertinentes si λ était trop grand.

Sélection Lasso bootstrappé randomisé (BRLS)

Pour régler ce problème du choix de λ, Meinshausen et Buehlmann 2010 ont proposé une généralisation de la procédure Lasso, appelée Lasso randomisé, où les covariables sont pénalisées par différentes valeurs aléatoirement choisies dans l'intervalle $[\lambda,\lambda/\alpha]$ où α est dans $]0,1[$. Le paramètre β est estimé sous la contrainte $\sum_{j=1}^{p}\left|\frac{\beta_j}{W_j}\right| \leq \lambda$. En pratique, les covariables $\{X_j : \ j = 1,\ldots,p\}$ sont pondérées par les $\{W_j : \ j = 1,\ldots,p\}$ aléatoirement générés avec $P(W_j = \alpha) = p_w$ et $P(W_j = 1) = 1 - p_w$ où $p_w \in]0,1[$. Les auteurs ne donnent pas d'indication sur le choix de p_w et dans la suite, nous avons fixé sa valeur à 0,5. Cette procédure, qui peut aussi être combinée avec des échantillons bootstrap comme dans le cas du Bolasso, est très simple à implémenter et les auteurs ont montré que choisir α dans l'intervalle $[0{,}2; 0{,}8]$ donne une sélection de variables stable quel que soit le choix du paramètre λ.

Nous avons appliqué ces deux dernières méthodes initialement développées dans le cadre de la régression linéaire au cas du modèle de Cox.

4.1.3 Méthodes de sélection de variables basées sur les arbres de survie

Les arbres de survie sont la généralisation aux données censurées des arbres de régression et de classification, popularisés par l'algorithme CART (*Classification and Regression Tree*) de Breiman 1984 qui est basé sur un partitionnement récursif binaire. C'est un processus itératif qui divise les données en deux sous-groupes (les nœuds fils) selon la valeur des prédicteurs. La règle de division maximise la différence entre les deux nœuds fils. Le processus continue jusqu'à ce que chaque nœud atteigne une taille minimale précisée par l'utilisateur ou soit homogène. On parle alors de nœud terminal. Pour contrôler la taille de l'arbre et éviter le sur-ajustement, une règle d'arrêt est utilisée pour élaguer les grands arbres contenant des nœuds terminaux « purs ». La méthode employée par CART est basée sur une mesure de coût-complexité. La complexité d'un arbre Θ est :

$$R_{cp}(\Theta) = R(\Theta) + cp\left|\tilde{\Theta}\right|, \tag{4.1}$$

où $R(\Theta)$ est l'erreur totale de mesure (somme des erreurs de mesure de tous les nœuds terminaux), $\left|\tilde{\Theta}\right|$ est le nombre de nœuds terminaux et cp est un paramètre de pénalité positif appelé paramètre de complexité. Une séquence d'arbres emboîtés est construite. L'arbre final est le plus petit arbre pour lequel $R_{cp}(\Theta)$ est minimal. Le choix du paramètre cp est réalisé par des techniques de validation croisée.

Diverses adaptations de la méthode CART aux données de survie ont été proposées dès les années 1980 (voir Bou-Hamad et al. 2011 pour une revue historique). Elles diffèrent par le critère de division et par les règles d'élagage. Gordon et Olshen 1985 utilisent la métrique de Wasserstein pour mesurer la distance entre deux estimateurs de Kaplan-Meier de la fonction de survie. Le critère de division choisit le prédicteur (et le seuil correspondant) qui maximise cette distance pour les deux nœuds créés. Pour élaguer l'arbre, ils généralisent la mesure de coût-complexité aux données censurées. D'autres auteurs utilisent la statistique du test du logrank comme critère de division (voir Ciampi et al. 1986 ; Segal 1988 ; LeBlanc et Crowley 1993). Comme alternative à l'élagage standard qui utilise les résidus de martingale du modèle à risques proportionnels comme erreurs de mesure, LeBlanc et Crowley 1993 ont proposé une mesure de

complexité basée sur la somme des statistiques de test du logrank (voir 1.3) des nœuds internes. D'autres auteurs suggèrent d'utiliser un critère de division basé sur la fonction de vraisemblance : R. Davis et Anderson 1989 supposent un modèle exponentiel à chaque nœud tandis que LeBlanc et Crowley 1992 font seulement l'hypothèse que les risques instantanés des deux nœuds fils sont proportionnels. Ces derniers auteurs utilisent pour l'estimation la vraisemblance complète ou bien la vraisemblance partielle du modèle de Cox. A noter qu'il n'est pas possible de résumer l'estimation de la variable réponse d'un nœud terminal par une unique valeur comme dans les arbres de régression et de classification. A la place, on trouve l'estimateur de Kaplan-Meier de la survie du nœud — ou un estimateur du risque cumulé — avec éventuellement une estimation du risque relatif de ce nœud par rapport à la population globale, sous l'hypothèse des risques proportionnels.

L'instabilité des arbres est bien connue. Elle peut provenir d'un sur-ajustement de l'arbre aux données, mais aussi du choix arbitraire du seuil pour dichotomiser les prédicteurs continus à chaque nœud (voir Dannegger 2000). Nous présentons deux remèdes possibles à cette instabilité : les forêts aléatoires de survie de Ishwaran, Kogalur et al. 2008 et la stabilisation bootstrap au niveau des nœuds de Dannegger 2000.

Les forêts aléatoires de survie (RSF)

Pour stabiliser les arbres obtenus par l'algorithme CART, Breiman 1996 a proposé la méthode du *bagging* (pour *bootstrap aggregating*), qui consiste à agréger des arbres construits sur des échantillons bootstrap pour obtenir un estimateur robuste. La technique du *bagging* a été adaptée au contexte de la survie par Efron 1981 et Akritas 1986. Un autre moyen d'améliorer la stabilité des arbres est le *boosting*, développé par Freund et Schapire 1999. Comme le *bagging*, le *boosting* consiste à agréger une famille d'arbres. Chaque arbre est construit de façon itérative à partir d'un échantillon pondéré (un individu mal classé gagne en poids tandis qu'un individu bien classé en perd) et évalué en fonction de sa capacité à classer les individus. Considéré comme plus efficace que le *bagging*, le *boosting* est cependant limité quand les données sont trop bruitées : l'algorithme donne un poids élevé aux données bruitées ce qui aboutit à un mauvais ajustement (voir Dietterich 1999).

Breiman 2001 a proposé une méthode de sélection de variables qui combine la méthode du *bagging* avec une sélection aléatoire d'un ensemble de covariables à chaque nœud de l'arbre. Cette méthodologie, appelée « forêts aléatoires », s'est révélée plus stable que les deux méthodes précédentes (Breiman 1996 ; Freund et Schapire 1999). Cette méthode a été adaptée au domaine de la survie par Ishwaran, Kogalur et al. 2008 et est appelée « forêts aléatoires de survie » (*random survival forests* - RSF). Des échantillons bootstrap sont tirés de l'échantillon original. On peut noter que chaque échantillon bootstrap exclut environ 37 % des individus, un ensemble appelé données *out-of-bag* (OOB). Pour chaque échantillon bootstrap b (b variant de 1 à $ntree$), un arbre de survie est construit : à chaque nœud de l'arbre, un sous-ensemble de covariables est sélectionné aléatoirement parmi l'ensemble des covariables. Le nœud est divisé en deux nœuds fils à partir des covariables sélectionnées en maximisant le critère de division. Le processus de division continue jusqu'à ce que chaque nœud terminal contienne un nombre minimum fixé d'événements non censurés avec des temps distincts. L'algorithme RSF calcule alors un estimateur global qui est la moyenne des fonctions de risque cumulées estimées pour chaque arbre par l'estimateur de Nelson-Aalen (voir 1.2). Pour chaque nœud terminal h, soient $t_{(1,h)} < \cdots < t_{(l,h)} < \cdots < t_{(n_h,h)}$ les temps distincts ordonnés des événements non censurés des individus du nœud terminal h, et soient $m_{(l,h)}$ et $r_{(l,h)}$ le nombre d'événements et d'individus à risque au temps $t_{(l,h)}$, pour $l = 1, \ldots, n_h$.

L'estimateur de Nelson-Aalen du risque cumulé au nœud h est

$$\widehat{H}_h(t) = \sum_{(l,h):t_{(l,h)}<t} \frac{m_{(l,h)}}{r_{(l,h)}}.$$

Chaque arbre fournit de tels estimateurs du risque cumulé pour chacun de ses nœuds terminaux. Notons $\widehat{H}_b(t|x)$ la fonction de risque cumulée conditionnelle à la covariable x estimée par l'arbre b. Pour déterminer l'estimation du risque cumulé d'un individu i de covariable X_i obtenue à partir de l'arbre b, il suffit de faire parcourir l'arbre au vecteur de covariables X_i. Cela nous amène à un unique nœud terminal h. Ainsi nous obtenons $\widehat{H}_b(t|X_i) = \widehat{H}_h(t)$ si $i \in h$. Soit $I_{i,b} = 1$ si i est un individu OOB pour l'arbre b et 0 sinon. L'estimateur OOB global du risque cumulé de l'individu i est alors :

$$\widehat{H}^*_e(t|X_i) = \frac{\sum_{b=1}^{ntree} I_{i,b}\widehat{H}_b(t|X_i)}{\sum_{b=1}^{ntree} I_{i,b}}. \tag{4.2}$$

Cet estimateur est obtenu en faisant la moyenne des estimateurs du risque cumulé obtenus à partir des échantillons bootstrap qui excluent l'individu i.

On peut classer les impacts des variables sélectionnées par la procédure RSF sur l'événement d'intérêt grâce à la mesure de l'importance (notée VIMP) de chaque covariable, qui est la différence entre l'erreur de prédiction obtenue avec la forêt originale et l'erreur de prédiction obtenue en utilisant des affectations aléatoires à chaque nœud où la covariable considérée est rencontrée (voir ISHWARAN, KOGALUR et al. 2008).

Arbre de survie avec sélection de la division par bootstrap (BNLS)

DANNEGGER 2000 a proposé une procédure de stabilisation par bootstrap au niveau de chaque nœud pour les arbres de survie. L'algorithme consiste, au niveau du nœud h, à bootstrapper des échantillons à partir de l'ensemble des données présentes à ce nœud et, pour chacun d'eux, la meilleure division est recherchée. La division qui apparaît le plus de fois est sélectionnée. Pour une variable quantitative, la valeur seuil b choisie dans l'ensemble des réalisations de la variable de division est la médiane de toutes les valeurs seuil proposées à chaque échantillon bootstrap. DANNEGGER 2000 ne propose pas de méthode de choix de seuil pour les variables qualitatives. Nous avons donc fait le choix suivant : nous affectons chaque modalité à la branche où elle est le plus souvent allouée par les échantillons bootstrap.

DANNEGGER 2000 a comparé cette méthode (pour 100 bootstraps à chaque nœud) à l'algorithme CART et à la méthode du *bagging* dans le cas de données simulées. Il a répété l'expérience 50 fois et a constaté que sa méthode réduit les erreurs de prédiction de 26% (valeur obtenue avec l'algorithme CART) à 6%. La méthode de *bagging* l'emporte avec une erreur de prédiction de 3%. Mais le modèle obtenu à l'issue du *bagging* n'est pas facile à utiliser et à interpréter pour les non statisticiens. DANNEGGER 2000 conseille donc d'utiliser sa technique couplée à la validation croisée afin de trouver la valeur optimale du paramètre de complexité *cp* qui minimise le taux d'erreur de prédiction.

4.1.4 Comparaison des méthodes sur deux jeux de données

Afin de comparer les différentes approches de sélection de variables, nous les avons appliquées à deux jeux de données réels : le premier est un jeu de données public sur

le cancer du sein et le deuxième est une base de données originale sur la fertilité. Les cinq procédures suivantes ont été comparées : la sélection *stepwise* de Cox bootstrappée (*Bootstrap Stepwise Selection* - BSS), la procédure de Cox avec Lasso bootstrappée (*Bootstrap Lasso Selection* - BLSS), la procédure de Cox avec Lasso randomisé bootstrappée (*Bootstrap Randomized Lasso Selection* - BRLS) avec trois différentes valeurs de α (0,2, 0,4 et 0,6), la procédure de DANNEGGER 2000 de sélection de la division de l'arbre à chaque nœud par bootstrap (*Bootstrap Node Level Selection* - BNLS) et la procédure des forêts aléatoires de survie (*Random Survival Forest* - RSF).

Logiciel

Les données ont été analysées à partir des méthodes implémentées dans le logiciel `R`. Nous avons adapté les techniques de bootstrap à l'algorithme *stepwise* dans la procédure `step` de `R` et à la pénalisation $\mathcal{L}_1$ à partir de la bibliothèque `penalized` de `R`. Nous avons également modifié la bibliothèque `rpart` de `R` afin d'introduire la sélection bootstrap de la division de l'arbre à chaque nœud dans la construction de l'arbre (procédure BNLS). A la place de la méthode *exp* qui maximise la vraisemblance exponentielle, nous avons utilisé comme critère de division du nœud la statistique de test du logrank, en suivant les recommandations faites par RADESPIEL-TROGER et al. 2006. Ils ont montré que l'erreur de prédiction est liée à l'instabilité de la sélection de la division à chaque nœud. Comparé à d'autres critères de division sur un jeu de données réel de survie, le test du logrank fournit la plus faible erreur de prédiction (évaluée par le score de Brier, voir GRAF et al. 1999). La bibliothèque `randomSurvivalForests` de `R` propose quatre règles de division différentes : la statistique de test usuelle du logrank, la règle de conservation des événements, la statistique du test du score du logrank de HOTHORN et LAUSEN 2003 et une approximation de la statistique de test du logrank, plus rapide à calculer (pour plus de détails, voir ISHWARAN et KOGALUR 2007). Nous avons décidé d'utiliser la règle par défaut, *i.e.* la statistique de test usuelle du logrank pour mettre les deux procédures basées sur les arbres dans les mêmes conditions. Pour ce qui est du nombre d'échantillons bootstrap, nous avons pris $N = 100$ pour les procédures basées sur le modèle de Cox, la valeur par défaut de $N = 1000$ pour RSF et $N = 1000$ pour BNLS.

Critères de comparaison

Nous avons comparé les cinq procédures sur la base de leur performance prédictive et des gènes qu'elles sélectionnent sur les échantillons bootstrappés.

La performance prédictive est mesurée par le taux d'erreur de prédiction, calculé sur un échantillon de validation après avoir ajusté le modèle sur un échantillon d'apprentissage. Le taux d'erreur de prédiction le plus couramment utilisé pour les modèles de survie est basé sur l'indice de concordance de HARRELL et C. E. DAVIS 1982, appelé C-index. Afin de calculer cet indice, les temps de survie observés et les valeurs prédites sont comparés : on considère uniquement les paires d'observations admissibles, qui sont toutes les paires possibles d'observations moins les paires où la plus courte durée de survie est censurée ainsi que les paires qui ont les deux durées de survie et les deux indicatrices d'événement égales. Pour chaque paire admissible, on compte 1 si les prédictions sont concordantes avec les observations, c'est-à-dire ici si la plus courte durée de survie correspond à la prédiction la plus pessimiste. On compte 0,5 en cas de prédictions identiques. Le C-index est la moyenne des valeurs obtenues sur l'ensemble des paires admissibles. Le taux d'erreur de prédiction est $1 - C$-index et appartient donc à [0,1]. On peut noter qu'une valeur de 0,5 indique que la méthode ne prédit pas mieux que le hasard.

Pour le modèle de Cox, la prédiction pour un individu i de covariable X_i est le prédicteur linéaire $\hat{\beta}' X_i$ [1]. L'individu i a un plus grand risque d'avoir l'événement d'intérêt que l'individu j de covariable X_j si $\hat{\beta}' X_i > \hat{\beta}' X_j$. Pour les arbres de survie, la prédiction pour l'individu i est basée sur l'estimateur $\hat{H}(t|X_i)$ du risque cumulé. Pour la procédure RSF, les prédictions sont basées sur l'estimateur global OOB $\hat{H}^*_e(t|X_i)$ du risque cumulé (cf. 4.2). Pour que la prédiction soit indépendante du temps d'évaluation, on somme les estimations du risque cumulé sur tous les temps de survie non censurés distincts $t^*_1, \ldots, t^*_k$. On considère que l'individu i a un plus grand risque d'avoir l'événement d'intérêt que l'individu j si

$$\sum_{\ell=1}^{k} \hat{H}^*_e(t^*_\ell|X_i) > \sum_{\ell=1}^{k} \hat{H}^*_e(t^*_\ell|X_j).$$

Afin d'obtenir un échantillon de taux d'erreur de prédiction pour chaque méthode, 30 échantillons d'apprentissage et de validation ont été construits. Nous avons calculé les moyennes et écarts-types de ces taux d'erreur et tracé des boîtes à moustaches afin d'étudier la variabilité des taux d'erreur associés à chaque procédure. Cela nous permet de mesurer la stabilité de la procédure en termes de taux d'erreur.

Application au jeu de données sur le cancer du sein

Description

Les données sur le cancer du sein proviennent d'une étude de VIJVER et al. 2002 sur la durée de survie sans métastases de 295 patientes atteintes d'un cancer du sein primaire. Le but de l'étude est de valider un classement en deux groupes de bon et mauvais pronostic basé sur le profil génomique des patientes [2], établi par une étude antérieure, la fameuse signature des 70 gènes d'Amsterdam de VEER et al. 2002. Nous avons restreint l'étude aux 144 patientes qui présentaient des ganglions lymphatiques infectés. Le jeu de données peut être téléchargé à partir de la bibliothèque `penalized` de `R`. Cinq facteurs de risque cliniques ainsi que les mesures d'expression génomique des 70 gènes pronostiques de l'apparition de métastases y sont enregistrés. Le taux de censure est de 66%. Les covariables cliniques sont plus précisément :

- `Diam` : le diamètre de la tumeur (deux modalités : $\leq$ 2cm ou $>$ 2cm),
- `N` : le nombre de ganglions lymphatiques infectés (deux modalités : $1-3$ ou ≥ 4),
- `ER` : la présence de récepteurs œstrogéniques,
- `Grade` : le stade de la tumeur (3 modalités ordonnées),
- `Age` : l'âge de la patiente à la date du diagnostic.

Résultats

Pour la procédure BSS, l'algorithme *stepwise* appliqué à la régression de Cox n'a pas convergé, sans doute à cause d'un trop grand nombre de covariables et à un nombre trop faible d'événements dans la base de données bootstrappée. Le paramètre de complexité *cp* de la procédure BNLS a été obtenu par une validation croisée où l'échantillon a été découpé en 10 parties : une valeur de $cp = 0{,}002$ semble optimale pour obtenir l'arbre le plus prédictif. La figure 4.1 présente les boîtes à moustaches des taux d'erreur de prédiction des différentes approches pour 30 échantillons d'apprentissage et de validation. Les résultats numériques sont résumés dans le tableau 4.1. Comme attendu, la procédure RSF donne les plus faibles taux d'erreur de prédiction, avec la plus petite variation et

1. $\widehat{\beta}$ désigne l'estimateur du p-vecteur des paramètres β du modèle de Cox 1.4, obtenu en maximisant la log-vraisemblance partielle 1.5.
2. Il s'agit donc d'une signature génomique, voir section 4.2.

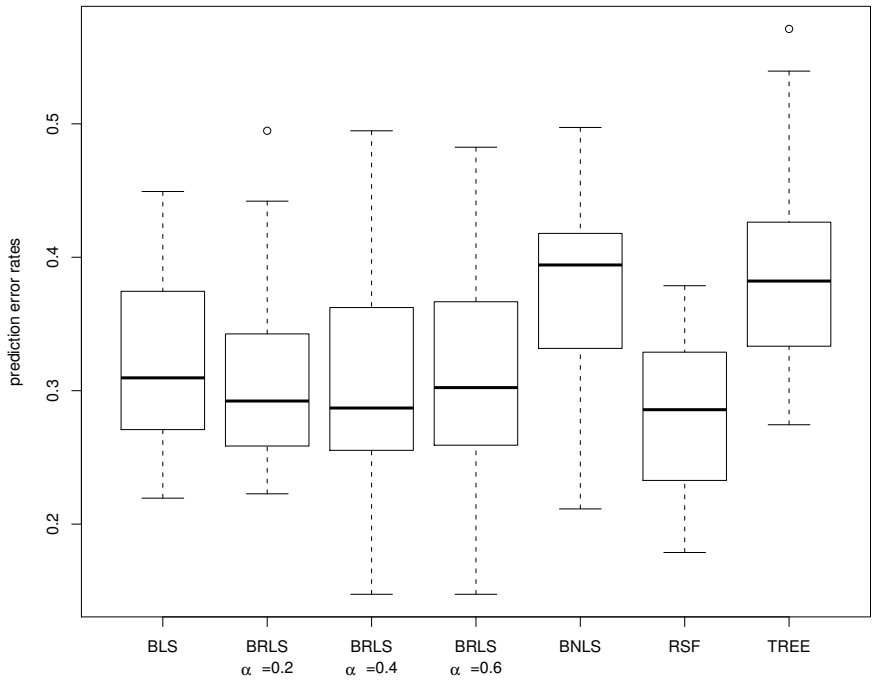


FIGURE 4.1 – Boîtes à moustaches des taux d'erreur de prédiction sur les données du cancer du sein pour les procédures *bootstrap lasso selection*(BLS), *bootstrap randomized lasso selection*(BRLS), *bootstrap node-level selection*(BNLS), *random survival forest*(RSF) et un arbre de survie (TREE).

un simple arbre de survie fournit les plus mauvais taux d'erreur de prédiction, ces taux n'étant que faiblement améliorés par la procédure BNLS. Quelle que soit la valeur de α, la procédure BRLS donne des taux d'erreur plus dispersés que ceux obtenus par la procédure BLS pour laquelle l'écart-type est de 0,06, similaire à celui de RSF. Toutefois, un bon compromis entre dispersion et moyenne du taux d'erreur de prédiction est observé pour la méthode BRLS avec $\alpha = 0{,}2$ ou $\alpha = 0{,}4$.

En ce qui concerne les covariables sélectionnées par les procédures basées sur le modèle de Cox, seules 23 des 75 covariables sont incluses dans les sélections faites par les procédures appliquées aux échantillons bootstrappés et aucune covariable clinique n'est retrouvée. Pour une valeur de κ égale à 0,2, 6 gènes sont sélectionnés par BLS et seulement 4 (inclus dans les 6 précédents) par BRLS quelle que soit la valeur de α. Pour les procédures RSF et BNLS, le premier gène discriminant est le même et il est parmi les quatre gènes retenus par BRLS. La méthode BNLS sélectionne également 4 autres covariables parmi les 6 variables les plus importantes de la procédure RSF, dont deux covariables cliniques (`N` et l'âge). Comparé à un simple arbre de survie, nous trouvons les mêmes premières variables discriminantes. Mais les autres covariables sélectionnées par l'arbre unique ne sont jamais retrouvées dans les autres procédures. Nous pouvons également observer qu'aucune variable clinique n'est incluse dans le simple arbre de survie contrairement aux procédures BNLS et RSF.

Application au jeu de données sur la fertilité

Les données sur la fertilité ont été obtenues à partir de 2138 couples ayant consulté

	Moyenne	Écart-type	Médiane
BLS	0,319	0,066	0,310
BRLS $\alpha = 0{,}2$	0,309	0,075	0,292
BRLS $\alpha = 0{,}4$	0,309	0,081	0,287
BRLS $\alpha = 0{,}6$	0,311	0,086	0,302
BNLS	0,376	0,073	0,394
RSF	0,279	0,062	0,286
TREE	0,389	0,074	0,382

TABLE 4.1 – Moyenne, écart-type et médiane des taux d'erreur de prédiction sur les données du cancer du sein pour les procédures *bootstrap lasso selection*(BLS), *bootstrap randomized lasso selection*(BRLS), *bootstrap node-level selection*(BNLS), *random survival forest*(RSF) et un arbre de survie (TREE).

pour infertilité masculine au cours de la période allant de 2000 à 2004 au Centre de Stérilité Masculine de Toulouse (CSM). Les patients ont été suivis depuis leur admission lors de la première consultation et durant le traitement par un andrologue spécialiste jusqu'à l'arrêt du traitement ou jusqu'à la naissance d'un enfant vivant (la durée de suivi maximale est de 9 ans). L'issue reproductive est représentée par la naissance d'un enfant vivant à partir d'une grossesse obtenue dans le cadre du suivi au CSM (grossesse après un traitement médical et/ou traitement chirurgical, ou par une technique d'assistance médicale à la procréation (AMP) ainsi qu'après une grossesse spontanée). Ici, l'événement d'intérêt considéré est donc la naissance d'un enfant vivant et les observations censurées à droite correspondent aux perdus de vue ou aux couples n'ayant pas eu d'enfant vivant à la fin du suivi. Nous avons travaillé sur le sous-ensemble des 1773 couples pour lesquels aucune des covariables incluses dans la base de données ne comportait de valeurs manquantes. 40% des couples ont réussi à avoir un enfant, ce qui conduit à un taux de censure de 60 %. Nous avons décidé, en accord avec les cliniciens, de garder 32 covariables, parmi lesquelles l'âge et les antécédents médicaux de l'homme, l'âge et différents facteurs liés à la reproduction pour la femme ainsi que des variables du couple, comme la durée d'infertilité, le type d'infécondité (primaire ou secondaire) et les antécédents d'AMP du couple.

Résultats

Les taux d'erreur de prédiction pour chacune des cinq procédures sont présentés dans la figure 4.2 et le tableau 4.2. Nous pouvons observer que les moyennes et les médianes obtenues sont plus grandes que celles obtenues pour les données sur le cancer du sein, reflétant la difficulté de prédire le délai jusqu'à la naissance. Toutefois, la variabilité des taux d'erreur est plus petite. Comme pour le jeu de données sur le cancer du sein, la méthode RSF donne le meilleur modèle prédictif. Toutefois, il apparaît que la dispersion dans les taux d'erreur de RSF est similaire aux autres procédures. De plus, nous pouvons noter que la procédure BNLS n'est pas meilleure qu'un simple arbre de survie et cette constatation est valable également en comparant la procédure BSS avec une simple régression de Cox par sélection *stepwise*. Ces résultats peuvent être expliqués par la taille de l'échantillon qui est suffisamment grande ici pour produire de bons taux d'erreur de prédiction sans avoir recours à des techniques de ré-échantillonnage. Concernant les procédures BLS et BRLS, elles montrent moins de variations dans leurs taux d'erreur, mais leurs moyennes et leurs médianes sont proches de 0,5, suggérant que les modèles obtenus ne font pas mieux que le hasard.

Si maintenant on compare les covariables sélectionnées par les différentes procédures,

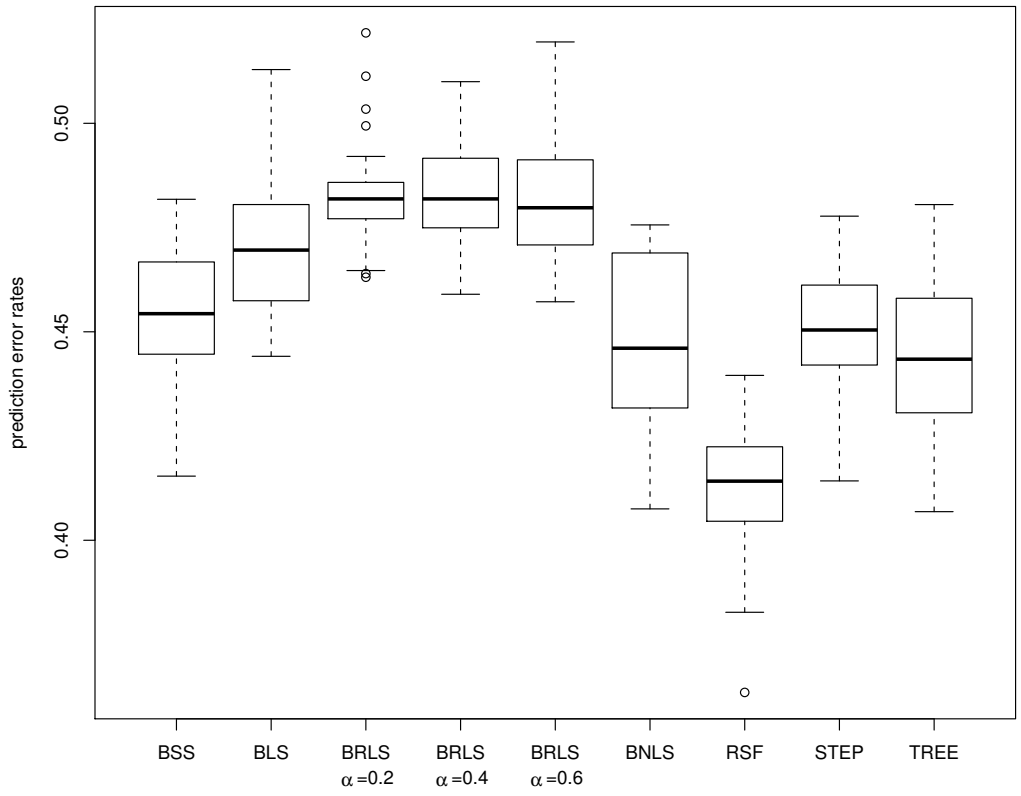


FIGURE 4.2 – Taux d'erreur de prédiction sur les données de fertilité pour les procédures *bootstrap stepwise selection*(BSS), *bootstrap lasso selection*(BLS), *bootstrap randomized lasso selection*(BRLS), *bootstrap node-level selection*(BNLS), *random survival forest*(RSF), *Cox stepwise selection*(STEP) et un arbre de survie (TREE).

	Moyenne	Écart-type	Médiane
BSS	0,453	0,017	0,454
BLS	0,471	0,017	0,470
BRLS $\alpha = 0{,}2$	0,483	0,013	0,482
BRLS $\alpha = 0{,}4$	0,484	0,013	0,482
BRLS $\alpha = 0{,}6$	0,482	0,015	0,480
BNLS	0,447	0,021	0,446
RSF	0,413	0,017	0,414
STEP	0,451	0,015	0,450
TREE	0,445	0,018	0,443

TABLE 4.2 – Moyenne, écart-type et médiane des taux d'erreur de prédiction sur les données de fertilité pour les procédures *bootstrap stepwise selection*(BSS), *bootstrap lasso selection*(BLS), *bootstrap randomized lasso selection*(BRLS), *bootstrap node-level selection*(BNLS), *random survival forest*(RSF), *Cox stepwise selection*(STEP) et un arbre de survie (TREE).

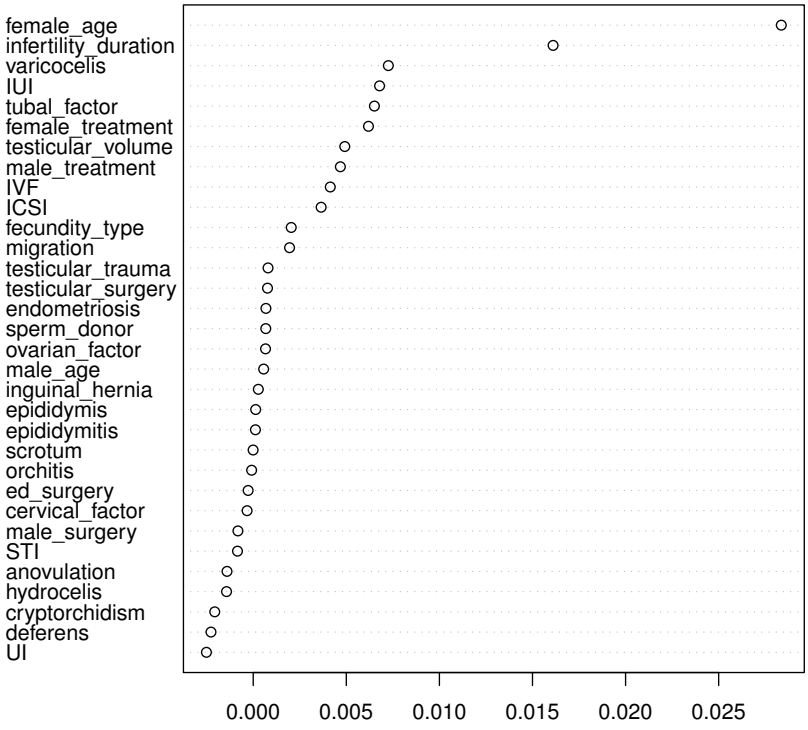


FIGURE 4.3 – Importance des variables selon la méthode RSF pour les données sur la fertilité.

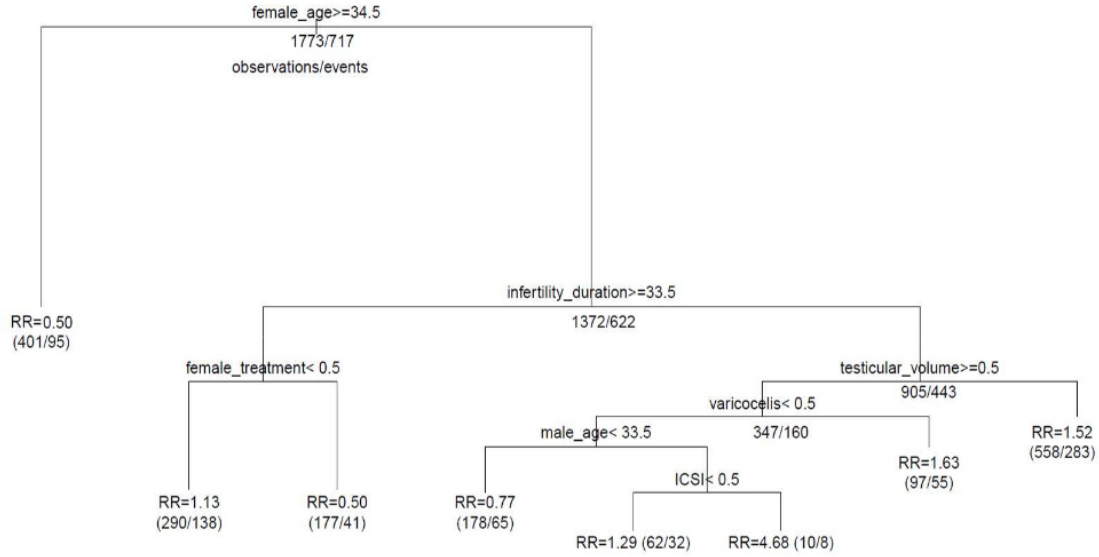


FIGURE 4.4 – Arbre final obtenu par la méthode BNLS sur les données de fertilité. La condition s'applique au nœud fils de gauche. RR est une estimation du risque relatif d'événement du nœud terminal par rapport au nœud initial.

nous pouvons constater que les quatre premières covariables sélectionnées ne diffèrent pas pour les procédures BLS et BRLS pour une valeur d'inclusion $\kappa = 0{,}3$: il s'agit de l'aspect des trompes, le recours à une insémination intra-utérine, le recours à des techniques d'AMP avec donneur de sperme et la durée d'infertilité du couple. Nous observons également que la procédure BSS inclut plus de variables que les procédures BLS et BRLS, conduisant à des taux d'erreur plus faibles. En ce qui concerne les procédures RSF et BNLS, on constate que les covariables sélectionnées sont sensiblement différentes de celles des autres procédures. Cela peut s'expliquer par le fait que ces procédures prennent en compte les interactions contrairement au modèle de Cox. Les variables retenues par la procédure BNLS (voir figure 4.4) sont retrouvées dans les variables de plus grande importance de RSF (voir figure 4.3) et les deux premières divisions de BNLS (âge de la femme avec un seuil à 34,5 ans et durée d'infertilité avec un seuil à 33,5 mois) sont aussi les deux variables qui ont les plus grandes importances avec RSF. Ces deux variables sont par ailleurs bien connues dans la littérature comme étant pronostiques de la naissance d'un enfant (voir ROCHEBROCHARD et THONNEAU 2002 ; MARQUARD et al. 2009 ; MALIZIA et al. 2009 ; DUCOT et al. 1988). Comparé à un simple arbre de survie, l'arbre construit par BNLS reste sensiblement le même ce qui peut là aussi s'expliquer par le grand nombre d'observations dans ce jeu de données.

4.1.5 Conclusion et perspectives

Pour ce qui est du jeu de données sur le cancer du sein, les taux d'erreur de prédiction sont plutôt corrects : la méthode basée sur la sélection Lasso randomisé combinée avec le ré-échantillonnage fournit des taux d'erreur de prédiction très similaires à ceux de RSF avec l'avantage d'un modèle facilement interprétable par les cliniciens. Par contre, la prédiction s'avère moins bonne pour le jeu de données sur la fertilité et les techniques de ré-échantillonnage n'ont pas montré d'amélioration notable.

Notre préconisation est que les procédures basées sur le modèle de Cox et les arbres de survie ont un rôle complémentaire à jouer afin d'identifier les covariables les plus pertinentes et fournir aux cliniciens un modèle stable et fiable.

La critique qui nous a été faite lors de la soumission de cet article est que nous n'avons pas comparé les performances des méthodes sur des données simulées. Nous comptons donc très prochainement resoumettre cet article en ajoutant la comparaison des différentes méthodes sur des données simulées comme cela a été fait dans l'article [14] (voir section 4.3.2), en générant en particulier des covariables qui n'ont aucun effet sur la survie.

4.2 Établissement d'une signature moléculaire en oncologie

Ces dernières années ont vu fleurir l'apparition de signatures génomiques en oncologie (voir par exemple VEER et al. 2002 et DOWSETT et al. 2013). A l'aube de la médecine personnalisée, cela permet d'améliorer la prise en charge des patients en prédisant l'issue du traitement (MICHIELS, TERNÈS et al. 2016). Une signature moléculaire (ou génomique) est simplement un score quantitatif ou qualitatif construit à partir de l'information génomique du patient afin d'apporter un bénéfice cliniquement perceptible en termes de pronostic ou d'indication thérapeutique.

Nous présentons dans cette section l'article [8], écrit en collaboration avec des cliniciens oncologues, qui ont fait appel à Marie Walschaerts et moi-même pour la construction d'une signature moléculaire pour séparer les patients en rémission d'un cancer du poumon en deux groupes de bon et mauvais pronostic.

Le cancer du poumon est le cancer le plus meurtrier au monde, avec environ un million de décès chaque année. La classification des cancers du poumon basée sur les stades cliniques s'avère insuffisante pour fournir un pronostic fiable de survie, en particulier pour les stades précoces. Il a par ailleurs été montré que des altérations dans le programme de réplication de l'ADN contribuent à la formation des tumeurs et que les cellules cancéreuses sont souvent exposées à un stress de réplication endogène. Des chercheurs en oncologie de Toulouse se sont intéressés à 77 gènes impliqués dans différents aspects de la réplication de l'ADN chromosomique. Ils ont comparé l'expression de ces gènes dans la tumeur et dans un tissu sain adjacent de 93 patients souffrant d'un cancer du poumon primaire à un stade précoce ou intermédiaire. 17 de ces 77 gènes se sont révélés avoir un niveau d'expression dans la tumeur significativement supérieur à celui dans le tissu sain. Ces niveaux d'expression ont été discrétisés pour chacun de ces gènes en trois classes sur la base des terciles. Parmi ces 17 gènes, 5 gènes (PLK1, CDC6, PLOQ, RAD51 et CLASPIN) ont un niveau d'expression significativement associé à la survie globale (l'événement est le décès toutes causes, le délai de survie étant mesuré depuis l'opération), à la survie sans maladie (l'événement est le décès toutes causes ou la rechute) et à la survie sans rechute (l'événement est la rechute). D'autre part, le niveau d'expression de ces gènes n'est significativement associé ni au traitement ni au stade du cancer. Enfin, excepté pour le gène RAD51, le rôle pronostique sur la survie globale de tous ces gènes persiste en ajustant sur l'âge, le sexe, le traitement, le stade du cancer et les marqueurs habituels de prolifération.

Avec Marie Walschaerts, nous avons alors établi une signature basée sur ces cinq gènes. Pour cela, nous avons ajusté aux données un modèle de Cox avec comme covariables les cinq gènes discrétisés en trois classes (sur la base des terciles de leur niveau d'expression). Le prédicteur linéaire du modèle de Cox $\widehat{\beta}'X$ (voir page 104), qui correspond donc à la somme des niveaux d'expression des cinq gènes pondérés par les coefficients correspondants du modèle de Cox, a été utilisé comme score de risque. Nous avons ensuite déterminé le seuil de ce score permettant d'obtenir deux groupes de bon et mauvais pronostic dont la différence de survie globale soit la plus significative possible par le test du logrank 1.3 (paquet `maxstat` de `R`). La figure 4.5 présente les estimations de Kaplan-Meier de la survie globale dans les deux groupes à faible et fort risque définis par cette signature. Le risque relatif de décès du groupe de mauvais pronostic par rapport au groupe de bon pronostic, établi à l'aide d'un modèle de Cox avec le groupe comme unique covariable, vaut 14,3 ($p < 0{,}001$). Cette signature devrait permettre de mieux prévoir l'espérance de vie d'un patient en rémission et d'adapter le traitement adjuvant prescrit à son groupe de risque.

4.3 Sélection de modèles de survie en présence de risques concurrents

4.3.1 Motivation

Comme rappelé dans la section 4.2, les signatures génomiques sont devenues ces dernières années un outil incontournable en oncologie pour aider au pronostic ou guider le choix du traitement. Malheureusement, plusieurs études ont montré que ces signatures n'étaient pas stables et dépendaient fortement de la méthode de régression utilisée (voir MICHIELS, KOSCIELNY et al. 2005 ; FERTÉ et al. 2013 ; EIN-DOR et al. 2005). D'autre part, les patients sont à risque de plusieurs événements concurrents et les signatures ont surtout été développées en considérant des événements composites, comme la survie sans rechute (voir KRAMAR et al. 2015). Il semble intéressant de pouvoir développer des signatures pour chaque type d'événement (voir DRUKKER et al. 2015 ; MITRA et al.

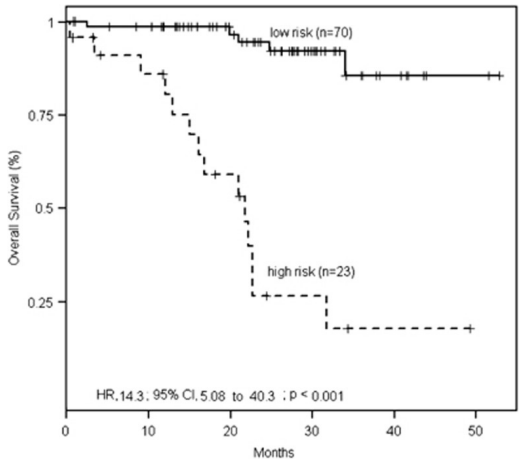


FIGURE 4.5 – Estimation de Kaplan-Meier de la survie globale dans les deux groupes à faible et fort risque construits à l'aide d'une signature génomique.

2014). A titre d'exemple, dans le cancer du sein, la récidive loco-régionale devient moins fréquente. Pour savoir quel traitement choisir dans ce cas-là, il est important de disposer d'une signature propre à ce risque de récidive, qui est en concurrence avec le risque de métastase à distance et le décès.

Les approches consistant à construire des signatures moléculaires dans le contexte des risques concurrents sont récentes et peu développées. Dans ce cadre, la modélisation la plus utilisée avec des données en grande dimension est l'approche suivante : un modèle de Cox avec pénalisation est ajusté sur l'événement d'intérêt, en considérant comme censurés les individus qui connaissent des événements concurrents (TAPAK et al. 2015). Outre le fait que cette procédure est discutable étant donné que les événements en concurrence sont rarement indépendants, une covariable qui réduit le risque d'un événement concurrent peut indirectement augmenter l'incidence cumulée de l'événement d'intérêt (voir DIGNAM et KOCHERGINSKY 2008). Rappelons que la fonction d'incidence cumulée (CIF) représente la probabilité d'observer l'événement en présence des événements concurrents et correspond à une sous-distribution. Dans le cadre des données en faible dimension, le modèle de FINE et GRAY 1999, qui est une extension du modèle de Cox, a été proposé pour modéliser le risque associé à la sous-distribution d'un événement. En cas de données en grande dimension, le modèle de Fine et Gray ne peut être utilisé pour des raisons techniques et il faut alors utiliser des approches moins classiques. Les méthodes basées sur les forêts aléatoires ont récemment été adaptées aux données de survie en présence de risques concurrents par ISHWARAN, GERDS et al. 2014. Une autre méthode proposée par BINDER et al. 2009 consiste à proposer une sélection de gènes basée sur le modèle de FINE et GRAY 1999 avec une approche utilisant le *boosting*. Ces deux méthodes sont maintenant implémentées dans le logiciel `R` mais à notre connaissance, elles n'ont jamais été comparées en termes de stabilité et de performance pronostique.

Dans l'article [14], nous comparons ces deux méthodes de sélection en utilisant un jeu de données publié concernant le cancer de la vessie ainsi que des données simulées. Une approche par ré-échantillonnage nous permet de comparer à la fois les gènes sélectionnés et la précision de la prédiction. Nous nous sommes également intéressés à l'effet de la taille des échantillons sur la performance de ces méthodes. Dans la section 4.3.2, nous présentons les jeux de données, les deux méthodes de sélection de variables et la

méthodologie de comparaison des deux méthodes. Les résultats sont exposés dans la section 4.3.3. Enfin, des conclusions et perspectives sont proposées dans la section 4.3.4.

4.3.2 Jeux de données et méthodes

Contexte

Nous nous plaçons dans le contexte des risques concurrents avec K événements en compétition, avec les notations déjà définies dans le chapitre précédent (voir page 76). Cela correspond au mécanisme de mélange censuré de la section 3.1.3. Nous observons pour chaque sujet le triplet (Y,Δ,X) où Y correspond au délai jusqu'au premier événement observé (de type k si $\Delta = k$) ou bien au délai de censure ($\Delta = 0$). X désigne un p-vecteur de covariables. Les quantités d'intérêt sont les fonctions d'incidence cumulée I_k pour l'événement de type k (voir leur définition en 3.34).

Pour modéliser la relation entre l'événement de type k et les covariables, nous utilisons le modèle de Fine et Gray 1999. Ce modèle stipule que le risque associé à la sous-distribution de l'événement d'intérêt (ici l'événement de type 1) vérifie le modèle suivant conditionnellement à la covariable X :

$$h_1(t \mid X) = h_{1,0}(t)e^{\beta' X}, \tag{4.3}$$

où $h_{1,0}$ est une fonction de risque de base (associé à une sous-distribution) non spécifiée et β le p-vecteur des paramètres.

Les données du cancer de la vessie

Nous nous sommes basés sur le jeu de données public GSE5479 — téléchargeable sur la plateforme GEO — correspondant à 1381 biopuces pour des prélèvements de 404 patients atteints d'un cancer de la vessie, qui sont les données utilisées par Dyrskjøt et al. 2005. Nous avons considéré uniquement les 301 patients pour lesquels on disposait d'un classifieur de progression et des covariables cliniques (âge, sexe, stade, grade du cancer et traitement). L'événement d'intérêt de notre étude est la progression du cancer ou le décès dû au cancer de la vessie : il a été observé pour 84 patients. L'événement concurrent est le décès dû à une autre cause ou à une cause inconnue, qui a été observé pour 33 patients. 184 patients ont été censurés à la fin de leur suivi. Les incidences cumulées à 5 ans pour l'événement d'intérêt et l'événement concurrent ont été respectivement estimées à 26,8 % et 11,2 % par l'estimateur de Kalbfleisch et Prentice 1980.

Les données simulées

Nous avons simulé quatre jeux de données de tailles différentes en utilisant l'algorithme utilisé par Binder et al. 2009 et Tapak et al. 2015. Pour chaque jeu de données, deux événements concurrents ont été générés selon le modèle exponentiel des causes spécifiques. Les délais de censure suivent une distribution uniforme avec un taux de censure d'environ 35 %. 1500 covariables normalement distribuées ont été générées. Seulement 16 d'entre elles étaient informatives et provenaient de trois groupes de covariables corrélées entre elles :

- le groupe 1 (corrélation de 0,5) contient 4 covariables qui augmentent les risques des deux types d'événement,
- le groupe 2 (corrélation de 0,35) contient 4 covariables qui augmentent le risque de l'événement de type 1 et diminuent le risque de l'événement de type 2,
- le groupe 3 (corrélation de 0,05) contient 4 covariables qui diminuent seulement le risque de l'événement de type 1 (pas d'effet sur l'événement de type 2) et 4 autres qui augmentent seulement le risque de l'événement de type 2 (pas d'effet sur l'événement de type 1).

Cela correspond donc aux vecteurs de paramètres β_1 pour l'événement de type 1 et β_2 pour l'événement de type 2 suivants :
$\beta_1 = (0{,}5; 0{,}5; 0{,}5; 0{,}5; 0{,}5; 0{,}5; 0{,}5; 0{,}5; -0{,}5; -0{,}5; -0{,}5; -0{,}5; 0; 0; 0; 0; 0)$,
$\beta_2 = (0{,}5; 0{,}5; 0{,}5; 0{,}5; -0{,}5; -0{,}5; -0{,}5; -0{,}5; 0; 0; 0; 0; 0; 0{,}5; 0{,}5; 0{,}5; 0{,}5)$.

Les méthodes de sélection

Les forêts aléatoires de survie

Les forêts aléatoires de survie (RSF), qui ont été présentées à la section 4.1.3, ont été adaptées pour prendre en compte des risques concurrents par Ishwaran, Gerds et al. 2014. Nous utilisons le paquet `randomForestSRC`. La règle utilisée pour scinder les nœuds est une statistique du test du logrank modifiée réalisant une bonne approximation de la statistique du test de Gray 1988. Pour sélectionner les variables, nous utilisons la "chasse" aux variables, qui est une méthode particulièrement adaptée aux données en grande dimension ; il s'agit d'une combinaison de deux techniques : l'importance des variables (VIMP) et la profondeur minimale.

L'approche par *boosting*

Nous avons appliqué l'approche par *boosting* pour risques concurrents proposée par Binder et al. 2009 pour le modèle de Fine et Gray 1999 modélisant le risque associé à la sous-distribution de notre événement d'intérêt. Cette approche sera notée CoxBoost dans la suite. L'approche par *boosting* est un algorithme de sélection pas-à-pas qui débute par l'ajustement du modèle sans covariables. A chaque pas, on met à jour l'estimation d'un seul paramètre en maximisant la log-vraisemblance partielle pénalisée du modèle. Une validation croisée par blocs de 10 a été effectuée pour trouver le nombre optimal de pas et la pénalité a été choisie grâce à la fonction `optimCoxBoostPenalty` du paquet `CoxBoost` du logiciel `R`.

Comparaison des performances des méthodes

Nous avons utilisé une stratégie de ré-échantillonnage, en tirant 100 échantillons d'apprentissage et 100 échantillons de validation à partir de l'échantillon initial. Nous avons fait varier la taille de l'échantillon d'apprentissage (1/3, 1/2 ou 2/3 de l'échantillon total). Pour chaque échantillon d'apprentissage, les deux méthodes de sélection ont été appliquées pour aboutir à un score de prédiction du risque et une classification en deux groupes de risque. Pour RSF, ce score correspond à l'estimation de la CIF à 5 ans calculée avec les individus *out of bag*. Pour Coxboost, le score est basé sur le prédicteur linéaire du modèle de Fine et Gray $\hat{\beta}' X_i$. Nous nous sommes basés sur les courbes ROC dépendantes du temps (Heagerty et al. 2000) pour dichotomiser ces scores afin d'obtenir un classifieur binaire de mauvais ou bon pronostic. Nous avons choisi le seuil de façon à obtenir à 5 ans une sensibilité d'au moins 90 % et la meilleure spécificité possible, ce qui correspond à une mauvaise classification pour moins de 10 % des patients qui ont rechuté pendant les 5 premières années (voir Zemmour et al. 2015).

Nous avons comparé le nombre de gènes sélectionnés par chaque méthode ainsi que leur fréquence parmi les 100 signatures. Nous avons ensuite appliqué le score de risque et la classification en deux groupes aux échantillons de validation pour évaluer les capacités prédictives des deux méthodes. Le risque relatif associé à la sous-distribution (SHR) a été estimé en ajustant un modèle de Fine et Gray avec le score de risque ou le groupe de risque comme prédicteur. Un indice de concordance qui généralise le C-index usuel (voir page 103) au cas des risques concurrents (noté C-index dans la suite, voir Wolbers et al. 2014) et le score de Brier intégré (Graf et al. 1999 ; Gerds et Schumacher 2006) ont alors été calculés à l'aide du paquet `pec` de `R` (Mogensen et al. 2012). L'indice de concordance fournit une mesure globale du pouvoir discriminatoire du modèle et a été

FIGURE 4.6 – Courbes des fonctions d'incidences cumulées obtenues pour les deux signatures identifiées par les méthodes RSF et CoxBoost.

calculé à la fois pour les scores et les groupes de risque. Les valeurs vont de 0,5 (pas de discrimination) à 1 (discrimination parfaite). Le score de Brier n'a été calculé que pour les scores de risque. Plus le score de Brier est bas et meilleure est la capacité prédictive du modèle. Nous avons également évalué les sensibilités et spécificités à 5 ans des groupes de risque à l'aide des courbes ROC.

4.3.3 Résultats

Illustration des méthodes sur les données du cancer de la vessie

Nous avons tiré un échantillon d'apprentissage de taille 100 du jeu de données du cancer de la vessie, ce qui correspond à 1/3 de l'échantillon initial. Nous observons 26 événements d'intérêt et 16 événements concurrents. RSF et CoxBoost identifient respectivement 68 et 4 gènes significatifs dont aucun en commun. Nous avons appliqué les scores de risque et les groupes à l'échantillon de validation. Pour les scores de risque, nous obtenons des C-index de 0,68 et 0,67 respectivement pour RSF et CoxBoost, et des scores de Brier de 0,19 et 0,21, donc des valeurs assez proches. La figure 4.6 présente les courbes des incidences cumulées obtenues pour l'événement d'intérêt pour les deux groupes de mauvais et bon pronostics identifiés par chacune des méthodes de sélection. Les incidences cumulées à 5 ans sont estimées par RSF à 11,7 % et 31,4 % dans les groupes de bon et mauvais pronostic respectivement (test de Gray : p = 0,0003, C-index = 0,58). En utilisant CoxBoost, les incidences cumulées à 5 ans sont estimées à 13 % et 33,8 % dans les deux groupes respectivement (test de Gray : p < 0,001, C-index =0,62). Les sensibilités et spécificités à 5 ans des signatures calculées sur l'échantillon de validation sont de 89,5 % et 32,8 % pour RSF et de 83,7 % et 39,1 % pour CoxBoost. En rajoutant les covariables cliniques dans un modèle multivarié, l'effet du groupe reste significatif pour les deux méthodes (sHR=2,83, IC à 95 % =[1,21 ;6,60], p=0,016 pour RSF ; sHR=3,41, IC à 95 % =[1,56 ;7,42], p=0,002 pour CoxBoost).

Comparaison sur les données avec ré-échantillonnage

Quelle que soit la taille de l'échantillon d'apprentissage, RSF sélectionne toujours plus de gènes que CoxBoost : entre 52 et 78 pour RSF et entre 0 et 43 pour CoxBoost. Pour CoxBoost, le nombre de gènes sélectionnés augmente avec la taille de l'échantillon

A B

FIGURE 4.7 – Boîtes à moustaches des C-index (figure A) et des scores de Brier (figure B) pour les scores de risques selon la méthode et la taille de l'échantillon d'apprentissage pour les données du cancer de la vessie.

d'apprentissage alors qu'il diminue pour RSF. Entre 2 et 5 gènes sont inclus dans plus de la moitié des échantillons pour CoxBoost contre aucun pour RSF. La stabilité des méthodes, mesurée par la fréquence de sélection des gènes parmi les 100 échantillons, augmente avec la taille de l'échantillon d'apprentissage. Très peu de gènes sélectionnés sont communs aux deux méthodes (1 en médiane). Quant aux C-index des scores de risque, à taille d'échantillon d'apprentissage fixée, ils sont assez similaires pour les deux méthodes, alors que les scores de Brier sont plus faibles pour RSF (voir figure 4.7). Comme attendu, la dichotomisation des scores de risque en deux groupes diminue la performance prédictive des méthodes (voir figure 4.8 A). CoxBoost montre des performances légèrement moins bonnes que RSF, qui a des intervalles inter-quartiles plus réduits.

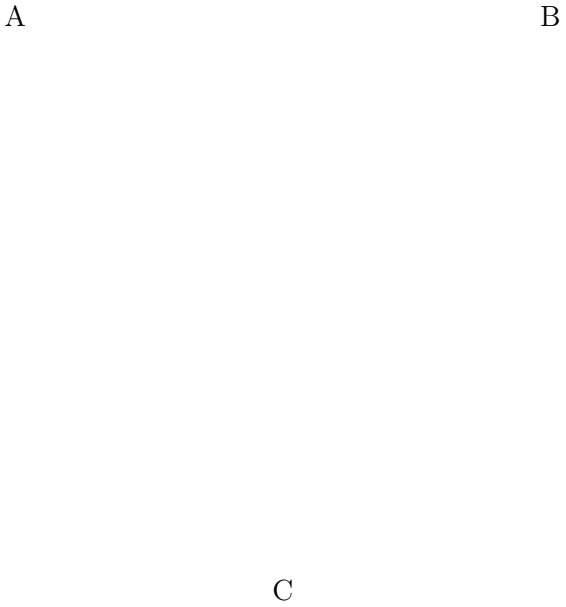
Résultats sur les données simulées

Les mêmes résultats sont obtenus en ce qui concerne l'effet de la taille de l'échantillon d'apprentissage : globalement, le nombre de gènes sélectionnés diminue pour RSF lorsque la taille augmente alors qu'il augmente pour CoxBoost. Cette tendance reste vraie pour ce qui concerne la sélection des gènes non informatifs. L'approche par *boosting* sélectionne plus que RSF les prédicteurs qui ont des effets contraires sur les deux risques ou un effet uniquement sur l'événement de type 1. On constate aussi que CoxBoost semble plus stable que RSF avec certaines covariables qui sont sélectionnées dans 100 % des échantillons. En ce qui concerne la performance prédictive, CoxBoost réussit mieux que RSF excepté pour la sensibilité. En outre, la taille de l'échantillon d'apprentissage a un effet plus important pour CoxBoost, en améliorant la prédiction alors que RSF est moins affecté. Pour RSF, cependant, c'est le nombre d'événements observés qui importe, en particulier pour les faibles tailles d'échantillons.

4.3.4 Conclusion et perspectives

La conclusion de [14] est que les performances prédictives des méthodes sont à peu près similaires et correctes lorsque la taille de l'échantillon et la proportion d'événements observés sont suffisantes mais que les signatures génomiques obtenues restent très

A



B

C

FIGURE 4.8 – Boîtes à moustaches des C-index (figure A), des risques relatifs associés aux sous-distributions (figure B) et des sensibilités (figure C) pour les signatures des deux méthodes selon la taille de l'échantillon d'apprentissage pour les données du cancer de la vessie.

instables. Une explication biologique peut être qu'il existe un grand nombre de gènes associés au même mécanisme d'action de la maladie étudiée, ce mécanisme étant corrélé au pronostic.

On peut noter que changer les options des fonctions de `R` n'a que peu d'incidence sur les résultats et nous pouvons donc garder les valeurs par défaut pour RSF. Cela concorde avec les conclusions de Díaz-Uriarte et De Andres 2006.

Comme indiqué dans la littérature (voir Dobbin et Simon 2010), nous avons retrouvé dans nos résultats le fait que la stabilité et le pouvoir prédictif des méthodes augmentent avec la taille de l'échantillon d'apprentissage. La proportion optimale recommandée par Dobbin et Simon 2010 varie entre 40 % et 80 % de la taille de l'échantillon global.

On pourrait par ailleurs envisager une réduction préalable de la dimension par une analyse en composantes principales ou une régression PLS. Une autre piste qui a été étudiée par Yuan et al. 2014 consiste à réduire le nombre de prédicteurs en ajustant aux données une forêt préliminaire qui calcule l'importance des variables. De nouvelles forêts sont ensuite construites avec les p prédicteurs qui ont la plus grande importance, en faisant varier p (de 10 à 50, par exemple). La forêt avec le taux d'erreur de prédiction le plus faible est ensuite sélectionnée.

Pour pouvoir comparer nos résultats avec ceux des autres études (voir Tapak et al. 2015 ; Binder et al. 2009), nous n'avons pas pris en compte les covariables cliniques dans notre sélection de variables. Cependant, l'inclusion des paramètres cliniques dans la construction du score pourrait sans doute permettre d'obtenir des signatures pronostiques plus précises.

Des méthodes de régression pénalisée comme la méthode Lasso ou Elastic net (voir Zou et Hastie 2005) ont été développées pour pallier les limitations du modèle de Cox. Dans le contexte des risques concurrents avec une approche cause-spécifique, Tapak et al. 2015 a comparé les performances prédictives des méthodes Lasso, Elastic net et CoxBoost. Elastic net a montré la meilleure performance. Cependant, les auteurs n'ont pas étudié la stabilité des gènes sélectionnés dans leurs simulations. Il serait intéressant d'adapter ces méthodes pénalisées au modèle de Fine et Gray de façon à pouvoir les comparer aux méthodes RSF et CoxBoost présentées ici. Dans le paquet `crrp`, Fu et al. 2016 propose un modèle de risques de sous-distributions proportionnels avec pénalisation mais son algorithme n'est pas directement applicable pour des données en grande dimension.

4.4 Références

Akritas, M. G. (1986). « Bootstrapping the Kaplan-Meier estimator ». In : *Journal of American Statistical Association* 81, p. 1032–1038.

Bach, F. (2009). « Model-Consistent Sparse Estimation through the Bootstrap ». Working paper or preprint. URL : https://hal.archives-ouvertes.fr/hal-00354771.

Binder, H., A. Allignol, M. Schumacher et J. Beyersmann (2009). « Boosting for high-dimensional time-to-event data with competing risks ». In : *Bioinformatics* 25.7, p. 890–896.

Bou-Hamad, I., D. Larocque et H. Ben-Ameur (2011). « A review of survival trees ». In : *Statistics Surveys* 5, p. 44–71. DOI : 10.1214/09-SS047.

Breiman, L. (1984). *Classification and Regression Trees*. Chapman et Hall/CRC.

— (1996). « Bagging predictors. » In : *Machine Learning* 24, p. 123–140.

— (2001). « Random Forests ». In : *Machine Learning* 45.1, p. 5–32.

Chen, C. H. et S. L. George (1985). « The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model ». In : *Statistics in Medicine* 4.1, p. 39–46.

Ciampi, A., J. Thiffault, J.-P. Nakache et B. Asselain (1986). « Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates ». In : *Computational Statistics and Data Analysis* 4, p. 185–204.

Dannegger, F. (2000). « Tree stability diagnostics and some remedies for instability ». In : *Statistics in Medicine* 19.4, p. 475–491. URL : https://epub.ub.uni-muenchen.de/1466/1/paper_72.pdf.

Davis, R. et J. Anderson (1989). « Exponential survival trees ». In : *Statistics in Medicine* 8, p. 947–962.

Díaz-Uriarte, R. et S. A. De Andres (2006). « Gene selection and classification of microarray data using random forest ». In : *BMC Bioinformatics* 7.1, p. 3. DOI : 10.1186/1471-2105-7-3.

Dietterich, T. (1999). *Ensemble methods in machine learning*. MCS '00 Proceedings of the First International Workshop on Multiple Classifier Systems. Springer-Verlag London, UK.

Dignam, J. J. et M. N. Kocherginsky (2008). « Choice and interpretation of statistical tests used when competing risks are present ». In : *Journal of Clinical Oncology* 26.24, p. 4027–4034. DOI : 10.1200/JCO.2007.12.9866.

Dobbin, K. K. et R. M. Simon (2010). « Optimally splitting cases for training and testing high dimensional classifiers ». In : *BMC Medical Genomics*.

Dowsett, M., I. Sestak, E. Lopez-Knowles, K. Sidhu, A. K. Dunbier, J. W. Cowens, S. Ferree, J. Storhoff, C. Schaper et J. Cuzick (2013). « Comparison of PAM50 Risk of Recurrence Score With Oncotype DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine Therapy ». In : *Journal of Clinical Oncology* 31.22, p. 2783–2790.

Drukker, C. A., S. G. Elias, M. Nijenhuis, J. Wesseling, H. Bartelink, P. Elkhuizen, B. Fowble, P. W. Whitworth, R. R. Patel, F. Snoo, L. van 't Veer, P. D. Beitsch et E. Rutgers (2015). « Erratum to: Gene expression profiling to predict the risk of locoregional recurrence in breast cancer: A pooled analysis ». In : *Breast Cancer Res Treat* 148, p. 599–613. DOI : 10.1007/s10549-014-3188-z.

Ducot, B., A. Spira, D. Feneux et P. Jouannet (1988). « Male factors and the likelihood of pregnancy in infertile couples. II. Study of clinical characteristics–practical consequences ». In : *Int J Androl* 11.5, p. 395–404.

Dyrskjøt, L., K. Zieger, M. Kruhøffer, T. Thykjaer, J. L. Jensen, H. Primdahl, N. Aziz, N. Marcussen, K. Jensen et T. F. Orntoft (2005). « A Molecular Signature in Superficial Bladder Carcinoma Predicts Clinical Outcome ». In : *Clinical cancer research* 11, p. 4029–36.

Efron, B. (1981). « Censored data and the bootstrap ». In : *Journal of American Statistical Association* 76, p. 312–319.

Ein-Dor, L., I. Kela, G. Getz, D. Givol et E. Domany (2005). « Outcome signature genes in breast cancer: is there a unique set? » In : *Bioinformatics* 21.2, p. 171–178. DOI : 10.1093/bioinformatics/bth469.

Ferté, C., A. D. Trister, E. Huang, B. M. Bot, J. Guinney, F. Commo, S. Sieberts, F. André, B. Besse, J. C. Soria et S .H. Friend (2013). « Impact of bioinformatic procedures in the development and translation of high-throughput molecular classifiers in oncology ». In : *Clinical Cancer Research* 19.16, p. 4315–4325.

Fine, J. P. et R. J. Gray (1999). « A Proportional Hazards Model for the Subdistribution of a Competing Risk ». In : *Journal of American Statistical Association* 94.446, p. 496–509. URL : http://www.jstor.org/stable/2670170.

Freedman, D. (1977). « A remark on the difference between sampling with and without replacement ». In : *Journal of American Statistical Association* 72.359, p. 681.

Freund, Y. et R. E. Schapire (1999). « A short introduction to boosting ». In : *Journal of Japanese Society for Artificial Intelligence* 14.5, p. 771–780.

Fu, Z., C. R. Parikh et B. Zhou (2016). « Penalized variable selection in competing risks regression ». In : *Lifetime Data Analysis* 23.03, p. 353–376.

Gerds, T. A. et M. Schumacher (2006). « Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times ». In : *Biometrical Journal* 48.6, p. 1029–1040. URL : http://dx.doi.org/10.1002/bimj.200610301.

Gey, S. et J. M. Poggi (2006). « Boosting and Instability for regression trees ». In : *Computational Statistics and Data Analysis* 50, p. 533–550.

Gordon, L. et R. A. Olshen (1985). « Tree-structured survival analysis ». In : *Cancer Treat Rep* 69.10, p. 1065–1069.

Graf, E., C. Schmoor, W. Sauerbrei et M. Schumacher (1999). « Assessment and comparison of prognostic classification schemes for survival data ». In : *Statistics in Medicine* 18.17-18, p. 2529–2545. URL : http://dx.doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5.

Gray, R. J. (1988). « A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk ». In : *The Annals of Statistics* 16.3, p. 1141–1154. URL : http://www.jstor.org/stable/2241622.

Harrell, F. E. et C. E. Davis (1982). « A new distribution-free quantile estimator ». In : *Biometrika* 69.3, p. 635–40.

Harrell, F. E., K. L. Lee, R. M. Califf, D. B. Pryor et R. A. Rosati (1984). « Regression modelling strategies for improved prognostic prediction ». In : *Statitics in Medicine* 3.2, p. 143–52.

Heagerty, P. J., T. Lumley et M. S. Pepe (2000). « Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker ». In : *Biometrics* 56.2, p. 337–344. URL : http://dx.doi.org/10.1111/j.0006-341X.2000.00337.x.

Hothorn, T. et B. Lausen (2003). « On the exact distribution of maximally selected rank statistics ». In : *Computational Statistics and Data Analysis* 43, p. 121–137.

Ishwaran, H., T. A. Gerds, U. Kogalur, R .D. Moore, S. Gange et B. M. Lau (2014). « Random survival forests for competing risks ». In : *Biostatistics* 15, p. 757–773. URL : https://www.researchgate.net/publication/261610216_Random_survival_forests_for_competing_risks.

Ishwaran, H. et U. B. Kogalur (2007). « Random Survival Forests for R ». In : *Rnews* 7/2, p. 25–31.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone et M. S. Lauer (2008). « Random Survival Forests ». In : *The Annals of Applied Statistics* 2.3, p. 841–860.

Kalbfleisch, J. D. et R. L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. John Wiley et Sons, Inc. ISBN : 9781118032985.

Kramar, A., S. Negrier, R. Sylvester, S. Joniau, P. Mulders, T. Powles, A. Bex, F. Bonnetain, A. Bossi, S. Bracarda, R. Bukowski, J. Catto, T. Choueiri, S. Crabb, T. Eisen, M. Demery, J. Fitzpatrick, V. Flamand, P. Goebell et T. Filleron (2015). « Guidelines for the definition of time-to-event

end points in renal cell cancer clinical trials: Results of the DATECAN project ». In : *Annals of Oncology* 26.12, p. 2392–2398.

LeBlanc, M. et J. Crowley (1992). « Relative risk trees for censored survival data ». In : *Biometrics* 48, p. 411–425.

— (1993). « Survival trees by goodness-of-split ». In : *Journal of the American Statistical Association* 88, p. 457–467. doi : 10.1080/01621459.1993.10476296.

Malizia, B. A., M. R. Hacker et A. S. Penzias (2009). « Cumulative live-birth rates after in vitro fertilization ». In : *New England Journal of Medicine* 360.3, p. 236–243.

Marquard, K., L. M. Westphal, A. A. Milki et R. B. Lathi (2009). « Etiology of recurrent pregnancy loss in women over the age of 35 years ». In : *Fertility and Sterility* 94.4, p. 1473–1477.

Meinshausen, N. et P. Buehlmann (2006). « High dimensional graphs and variable selection with the Lasso ». In : *Annals of Statistics* 34.3, p. 1436–1462.

— (2010). « Stability selection ». In : *Journal of the Royal Statistical Society: Series B* 72.4, p. 417–473.

Michiels, S., S. Koscielny et C. Hill (2005). « Prediction of Cancer Outcome With Microarrays: A Multiple Random Validation Strategy ». In : *The Lancet* 365, p. 488–492.

Michiels, S., N. Ternès et F. Rotolo (2016). « Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice ». In : *Annals of Oncology* 27.12, p. 2160–2167.

Mitra, A. P., L. L. Lam, M. Ghadessi, N. Erho, I. Vergara, M. Alshalalfa, C. Buerki, Z. Haddad, T. Sierocinski, T. Triche, E. Skinner, E. Davicioni, S. Daneshmand et P. Black (2014). « Discovery and Validation of Novel Expression Signature for Postcystectomy Recurrence in High-Risk Bladder Cancer ». In : *Journal of the National Cancer Institute* 106.11.

Mogensen, U. B., H. Ishwaran et T. A. Gerds (2012). « Evaluating Random Forests for Survival Analysis Using Prediction Error Curves ». In : *Journal of Statistical Software, Articles* 50.11, p. 1–23. url : https://www.jstatsoft.org/v050/i11.

Radespiel-Troger, M., O. Gefeller, T. Rabenstein et T. Hothorn (2006). « Association between split selection instability and predictive error in survival trees ». In : *Methods Inf Med* 45.5, p. 548–556.

Rochebrochard, E. de la et P. Thonneau (2002). « Paternal age and maternal age are risk factors for miscarriage; results of a multicentre European study ». In : *Human Reproduction* 17.6, p. 1649–1656.

Ruey-Hsia, L. (2001). « Instability of decision tree classification algorithms ». Thèse de doct. University of Illinois, Urbana-Champaign.

Sauerbrei, W. et M. Schumacher (1992). « A bootstrap resampling procedure for model building: application to the Cox regression model ». In : *Statistics in Medicine* 11.16, p. 2093–2109.

Segal, M.R. (1988). « Regression trees for censored data ». In : *Biometrics* 44, p. 35–48.

Tapak, L., M. Saidijam, M. Sadeghifar, J. Poorolajal et H. Mahjub (2015). « Competing Risks Data Analysis with High-dimensional Covariates: An Application in Bladder Cancer ». In : *Genomics, Proteomics and Bioinformatics* 13.3, p. 169–176.

Tibshirani, R. (1997). « The lasso method for variable selection in the Cox model ». In : *Statistics in Medicine* 16.4, p. 385–395.

Veer, L. J. van't, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards et S. H. Friend

(2002). « Gene expression profiling predicts clinical outcome of breast cancer ». In : *Nature* 415.6871, p. 530–536.

Vijver, M.J. van de et al. (2002). « A gene-expression signature as a predictor of survival in breast cancer ». In : *New England Journal of Medicine* 347.25, p. 1999–2009.

Wolbers, M., P. Blanche, M. T. Koller, J. C. Witteman et T. A. Gerds (2014). « Concordance for prognostic models with competing risks ». In : *Biostatistics* 15.3, p. 526–539. URL : http://dx.doi.org/10.1093/biostatistics/kxt059.

Yuan, Y., E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao, L. Han, X. Huang, M.S. Lawrence, J .N. Weinstein, J. M. Stuart, G. B. Mills, L. A. Garraway, A. A. Margolin, G. Getz et H. Liang (2014). « Assessing the clinical utility of cancer genomic and proteomic data across tumor types ». In : *Nature Biotechnology* 32.7, p. 644–652. DOI : 10.1038/nbt.2940.

Zemmour, C., F. Bertucci, P. Finetti, B. Chetrit, D. Birnbaum, T. Filleron et J.-M. Boher (2015). « Prediction of early breast cancer metastasis from DNA microarray data using high-dimensional cox regression models ». In : *Cancer Inform.* 14.Suppl 2, p. 129–138. URL : https://www.ncbi.nlm.nih.gov/pubmed/25983547.

Zou, H. et T. Hastie (2005). « Regularization and Variable Selection via the Elastic Net ». In : *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2, p. 301–320. URL : http://www.jstor.org/stable/3647580.

Conclusion générale

Plus de 25 ans après ma première rencontre avec les données censurées et les applications médicales, ce domaine m'intéresse toujours autant et je compte donc poursuivre mes activités de recherche dans ce même contexte.

Dans le domaine des sondages, très peu d'estimateurs ont été développés dans le cas où la variable d'intérêt est censurée à droite. Avec ma collègue Sandrine Casanova, nous avons donc l'intention de continuer à développer des outils méthodologiques adaptés à la censure à droite dans ce cadre. Plus précisément, nous pensons, après une approche basée sur un modèle, nous intéresser à l'approche intermédiaire dite « assistée par un modèle » où les poids de sondage sont également pris en compte.

En ce qui concerne la sélection de covariables dans les modèles de survie en grande dimension, cet axe de recherche est toujours au centre de mes préoccupations actuelles. La prépublication [19] va être enrichie de simulations et resoumise. D'autre part, après les co-encadrements des thèses de doctorat de Serge Somda et de Bastien Cabarrou (dont la soutenance est prévue pour fin 2018), Thomas Filleron vient de me proposer de codiriger avec lui le doctorat de Julia Gilhodes, biostatisticienne à l'Institut Claudius Regaud, sur le thème de la méthodologie biostatistique pour le développement de signatures génomiques en cancérologie. A l'ère de la médecine personnalisée et du séquençage très haut débit, le graal de l'oncologue moderne est l'individualisation de la prise en charge thérapeutique des patients. Dans ce but, depuis près de 15 ans, des signatures génomiques, visant à mieux stratifier les patients selon leur pronostic et leurs bénéfices potentiels à recevoir un traitement, ont été publiées dans différentes indications en cancérologie. Cependant, le nombre de signatures ayant prouvé leur utilité clinique reste très faible. De plus, l'engouement pour ces signatures génomiques a été tempéré par différentes publications. En effet, comme nous l'avons vu dans le chapitre précédent, les signatures manquent de stabilité. D'autre part, il a aussi été montré de façon surprenante que des signatures génomiques qui n'étaient pas reliées à la cancérologie, par exemple la résilience vis-à-vis de la défaite sociale chez la souris ou la signature du rire post-prandial, avaient également un impact pronostique sur le cancer. L'objectif de la thèse de Julia Gilhodes ne sera pas de remettre en question les signatures génomiques mais, à partir de données publiées et simulées, d'évaluer différentes approches de sélection et de combinaison de scores cliniques et génomiques afin de proposer des solutions pour

pallier leurs limites (nombre de gènes sélectionnés trop important, manque de stabilité dans la sélection, etc.).

Enfin, j'ai toujours été très intéressée par l'interdisciplinarité et outre les applications biomédicales, un domaine qui m'intéresse beaucoup est la psychologie. On rencontre dans ce domaine-là aussi fréquemment des données censurées à droite, en particulier quand on évalue des temps d'apprentissage. J'espère pouvoir profiter de la proximité physique du groupe IAST (Institute for Advanced Study in Toulouse) qui compte plusieurs chercheurs en psychologie pour développer des collaborations fructueuses dans ce domaine.

A — Curriculum Vitae

Née le 18 mars 1968, de nationalité française, mariée, 3 enfants

Laboratoires d'affectation pour la recherche

- Toulouse School of Economics (TSE-R)
- Membre associé de l'Institut de Mathématiques de Toulouse

A.1 Formation et diplômes

1991 Diplôme d'Ingénieur Civil des Mines (École des Mines de Saint-Etienne), mention Très bien

1992 DEA de Statistique et Santé, Université Paris XI (major)

1994 DU de l'Institut de Formation Supérieure Biomédicale, Université Paris XI

1995 Doctorat de l'Université Paris XI, spécialité *Biostatistique*, titre : « Modèles et tests pour l'analyse statistique des données censurées multivariées », directeur de recherche : J. Lellouch, INSERM U. 169, mention Très Honorable avec Félicitations

A.2 Parcours professionnel

1992–1995 Allocataire de Recherche et Moniteur de l'Enseignement Supérieur à l'Université Paris XI, en maîtrise de Sciences Biologiques et Médicales

1995–1996 Attachée Temporaire d'Enseignement et de Recherche à l'IUT de Paris V dans les départements Statistique et Traitement Informatique des Données et Informatique

Depuis septembre 1996 Maître de Conférences à l'Université Toulouse 1 Capitole (Hors Classe depuis septembre 2013)

A.3 Encadrements

A.3.1 Encadrement de stages, projets et mémoires

Septembre 2011–Janvier 2012 co-encadrement (avec Thomas Filleron, Institut Claudius Regaud) du projet de 5ème année de **Agnès Solomiac** et

Charlène Gaston (INSA filière GMM MMS) : « Comparison of confidence intervals for the survival function : application to cancer data ».

Avril–Septembre 2012 co-encadrement (avec Thomas Filleron, Institut Claudius Regaud) du stage de **Serge Somda** (M2 Sciences, Technologie, Santé mention Santé Publique, Spécialité Biostatistique, ISPED, Université Bordeaux Ségalen) : « Individualisation du suivi en fonction de facteurs pronostiques et du type de rechute ».

Février–Août 2014 co-encadrement (avec Thomas Filleron, Institut Claudius Regaud) du stage de **Soufiane Ajana** (M2 Recherche Épidémiologie Clinique, Université Toulouse 3 Paul Sabatier) : « Prédiction de la survie en cancérologie : comparaison de différentes approches ».

Mai–Juin 2015 encadrement du mémoire de recherche de **Quentin Vialle** (M1 Économie et Statistiques, Université Toulouse 1 Capitole) : « Robustness to non-normality in one-way ANOVA and alternative tests ».

Depuis 2012 encadrement des mémoires de M2 Statistique et Économétrie en formation à distance pour plusieurs étudiants chaque année.

A.3.2 Activités d'encadrement doctoral

Participation avec Christine Thomas à l'encadrement de la 3ème partie du doctorat en Mathématiques Appliquées de **Sandrine Casanova**, soutenu le 20 janvier 2000 ; titre : « Estimation non paramétrique des quantiles conditionnels », Université Toulouse 1 Capitole.

Participation avec Philippe Jeannin à l'encadrement de la 3ème partie du doctorat en Sciences Économiques de **Ibtissem Agueb**, soutenu le 1er décembre 2005 ; titre : « Économie du patrimoine immobilier des ménages en France », Université Toulouse 1 Capitole.

Participation avec Philippe Besse et Patrick Thonneau à l'encadrement de la 3ème partie du doctorat en Biostatistique - Epidémiologie de **Marie Walschaerts**, soutenu le 30 juin 2011 ; titre : « La santé reproductive de l'homme : épidémiologie et statistique », Université Toulouse 3 Paul Sabatier.

Co-encadrement avec Thomas Filleron (Institut Claudius Regaud) du doctorat en Mathématiques Appliquées de **Serge Somda**, soutenu le 15 septembre 2015 ; titre : « Individualisation du suivi post-thérapeutique des patients traités du cancer en fonction des facteurs pronostiques et du type de rechute », Université Toulouse 1 Capitole.

Co-encadrement avec Thomas Filleron (Institut Claudius Regaud) du doctoraten Mathématiques Appliquées de **Bastien Cabarrou** depuis janvier 2016 ; titre : « Prise en compte de l'hétérogénéité de la population âgée dans le schéma des essais cliniques de Phase I et II en oncogériatrie », Université Toulouse 1 Capitole.

Participation avec Pascal Maussion, Antoine Picot et Anne Ruiz-Gazen à l'encadrement de la 2ème partie du doctorat en Génie Electrique de **Farah Salameh**, soutenu le 7 novembre 2016 ; titre : « Méthodes de modélisation statistique de la durée de vie de composants en génie électrique », ENSEEIHT, INP de Toulouse.

Co-encadrement avec Thomas Filleron (Institut Claudius Regaud) du doctorat en Mathématiques Appliquées de **Julia Gilhodes** depuis janvier 2018 ;

titre : « Méthodologie biostatistique pour le développement de signatures génomiques en cancérologie », Université Toulouse 1 Capitole.

A.4 Participation à des jurys de thèse

Membre du jury de thèse de **Sandrine Casanova** (20 janvier 2000), Université Toulouse 1 Capitole, « Estimation non paramétrique des quantiles conditionnels ».

Membre du jury de thèse de **Karine Lhéritier** (28 février 2002), Université Montpellier 1, « Survie et risques compétitifs : Applications aux blessés médullaires tétraplégiques français ».

Membre du jury de thèse de **Ibtissem Agueb** (1er décembre 2005), Université Toulouse 1 Capitole, « Économie du patrimoine immobilier des ménages en France ».

Membre du jury de thèse de **Marie Walschaerts** (30 juin 2011), Université Toulouse 3 Paul Sabatier, « La santé reproductive de l'homme : épidémiologie et statistique ».

Membre du jury de thèse de **Serge Somda** (15 septembre 2015), Université Toulouse 1 Capitole, « Individualisation du suivi post-thérapeutique des patients traités du cancer en fonction des facteurs pronostiques et du type de rechute ».

Membre du jury de thèse de **Farah Salameh** (7 novembre 2016), INP Toulouse, « Méthodes de modélisation statistique de la durée de vie de composants en génie électrique ».

A.5 Contrats de recherche

Contrat DAER/9408140 pour le Conseil Régional de Midi-Pyrénées (1996–1998) : « Disparités démographiques et économiques dans la région Midi-Pyrénées » (en collaboration avec Y. Aragon, C. Diack, D. Haughton, J. Haughton, M. Johnson, E. Malin, A. Ruiz-Gazen et C. Thomas-Agnan).

Convention avec l'URCAM Midi-Pyrénées (1999–2003) : études d'économie de la santé (en collaboration avec I. Dubec et A. Ruiz-Gazen).

Aide à Projet Nouveau du CNRS (1999–2001) en tant que membre de l'équipe Économie et Statistique de la Santé du GREMAQ.

Contrat DAER/99008446 pour le Conseil Régional de Midi-Pyrénées (2000–2002) : « Modélisation, simulation et analyse des dynamiques spatiales de l'économie » (en collaboration avec Y. Aragon, O. Perrin, A. Ruiz-Gazen, C. Thomas-Agnan).

Appel à projets 2012 Soutien à la recherche mathématique et statistique appliquée à la cancérologie de l'Institut de Recherche en Santé Publique (2013–2015), responsable de l'équipe 3 : « Un nouvel algorithme de modélisation pour données de survie adapté à la recherche de traitements individualisés des cancers » (coordinateur : Jean-Marie Boher).

Appel à projets 2013 Transversalité de l'IDEX de l'Université de Toulouse (2014–2016) : « LIFESPAN : statistical methods for prognosis from accelerated to nominal regime » (en collaboration avec A. Ruiz-Gazen et des enseignants-chercheurs de l'ENSEEIHT).

A.6 Responsabilités collectives et activités d'animation scientifique

Membre de la Commission de spécialistes section 26, de l'Université Toulouse 1 Capitole de 1997 à 2008 (membre élu assesseur de 1997 à 2000).

Membre de la Commission de spécialistes section 26, de l'Université Toulouse 3 Paul Sabatier de 2002 à 2005.

Membre élu du Conseil de l'UFR **de Sciences Économiques** de l'Université Toulouse 1 Capitole de 1998 à 2002 et de 2005 à 2008.

Membre élu du Conseil d'Administration de l'Université Toulouse 1 Capitole de 2001 à 2004.

Organisatrice du séminaire de Statistique du GREMAQ pour l'année 2005–2006 et l'année 2012–2013 puis de TSE-R pour l'année 2016–2017.

Membre du comité d'organisation du 5ème colloque *Statistical Methods in Biopharmacy*, SFdS et EFSPI, 26 et 27 septembre 2005, Paris.

Membre du comité d'organisation des XXXV*èmes Journées de Statistique*, SFdS, mai 2013, Toulouse.

Membre élu du bureau du Département de Mathématiques de l'Université Toulouse 1 Capitole depuis 2014.

Rapports d'expertise scientifique pour les revues *Computational Statistics and Data Analysis*, *Journal de la Société Française de statistique*, *Journal of Statistical Planning and Inference*.

A.7 Conférences invitées

« Samples comparisons for censored multivariate censored data », 1994, atelier de formation INSERM sur l'analyse des données corrélées dans la recherche biomédicale, organisé par D. Clayton et S. Richardson, Le Vésinet (en collaboration avec T. Moreau et J. Lellouch).

A.8 Communications à des congrès avec comité de lecture sans publication des actes

« Une nouvelle famille de tests de rang pour la comparaison de deux distributions de survie bivariées en présence de censure », 1993, XXV$^{\text{èmes}}$ Journées de Statistique, Vannes (en collaboration avec T. Moreau et J. Lellouch).

« Un modèle de régression semi-paramétrique pour l'analyse de durées censurées multivariées », 1996, XVII$^{\text{ème}}$ Rencontre Franco-Belge de Statisticiens, Marne-la-Vallée (en collaboration avec T. Moreau et J. Lellouch).

« Quantiles de régression avec variable dépendante censurée », 1998, XXX$^{\text{èmes}}$ Journées de Statistique, Rennes (en collaboration avec S. Poiraud-Casanova et C. Thomas-Agnan).

« Estimation non paramétrique de quantiles conditionnels avec variable réponse censurée », 1999, XXXI$^{\text{èmes}}$ Journées de Statistique, Grenoble (en collaboration avec S. Poiraud-Casanova et C. Thomas-Agnan).

« Estimation non paramétrique de quantiles conditionnels avec variable réponse censurée », 2000, XXXII$^{\text{èmes}}$ Journées de Statistique, Fès (Maroc) (en collaboration avec S. Poiraud-Casanova et C. Thomas-Agnan).

« Estimation des moindres carrés asymétrique et application à une mesure de performance des médecins », 2001, XXXIII^{èmes} Journées de Statistique, Nantes (en collaboration avec Y. Aragon, R. Chambers et S. Poiraud-Casanova).

« Mesure d'efficacité d'une activité médicale par régression expectile », 2002, XXXIV^{èmes} Journées de Statistique, Bruxelles (en collaboration avec Y. Aragon et S. Poiraud-Casanova).

« Utilisation d'une régression expectile pour identifier des points non centraux », 2002, XXXIV^{èmes} Journées de Statistique, Bruxelles (en collaboration avec Y. Aragon, R. Chambers et S. Poiraud-Casanova).

« Estimation non paramétrique pour un processus de renouvellement censuré à droite – applications biomédicales », 2002, XXXIV^{èmes} Journées de Statistique, Bruxelles (en collaboration avec G. Derzko et P. Deheuvels).

« Determinants of physicians' fees : a probit model with spatial dependencies », 2003, 2^{nd} Spatial Econometrics Workshop, Dijon (en collaboration avec I. Dubec, J. Le Sage et A. Ruiz-Gazen).

« Inférence non paramétrique pour des événements répétés tronqués et censurés aléatoirement à droite – applications biomédicales », 2003, XXXV^{èmes} Journées de Statistique, Lyon (en collaboration avec G. Derzko).

« Estimation non paramétrique pour les systèmes multi-états : applications biomédicales », 2004, XXXVI^{èmes} Journées de Statistique, Montpellier (en collaboration avec G. Derzko).

« Estimation dans un modèle de Cox stratifié avec indicateurs de strates manquants », 2007, XXXIX^{èmes} Journées de Statistique, Angers (en collaboration avec J.-F. Dupuy).

« Using conditional quantiles to estimate the cumulative distribution function of a censored variable in a small area », 2008, XXXX^{èmes} Journées de Statistique, Ottawa (en collaboration avec S. Casanova).

« Estimation sur petits domaines de la fonction de répartition d'une variable censurée à l'aide de quantiles conditionnels », 2009, XXXXI^{èmes} Journées de Statistique, Bordeaux (en collaboration avec S. Casanova).

« Propriétés asymptotiques d'estimateurs non paramétriques model-based de la fonction de répartition sur un petit domaine », 2010, XXXXII^{èmes} Journées de Statistique, Marseille (en collaboration avec S. Casanova).

« Estimateurs model-based non paramétriques model-based de la fonction de répartition d'une variable censurée à droite sur petits domaines », 2012, 7ème Colloque Francophone sur les Sondages, Rennes (en collaboration avec S. Casanova).

« Estimation non paramétrique de la fonction de répartition d'une variable censurée à droite », 2014, 8ème Colloque Francophone sur les Sondages, Dijon (en collaboration avec S. Casanova).

« Estimation non paramétrique de la fonction de répartition d'une variable censurée à droite sur petits domaines », 2016, XXXXVIII^{èmes} Journées de Statistique, Montpellier (en collaboration avec S. Casanova).

A.9 Invitations à des séminaires

« Regression models and tests for the analysis of multivariate censored data », 1995, séminaire commun de statistique et d'économétrie, Center for Opera-

tions Research and Econometrics et Institut de Statistique, Université Catholique de Louvain, Belgique.

« Estimation non paramétrique de la fonction de répartition conditionnelle et de ses quantiles avec variable réponse censurée », 1999, Séminaire Agropolis, Montpellier.

« Estimation non paramétrique de la fonction de répartition conditionnelle et de ses quantiles avec variable réponse censurée », 2002, Séminaire de Biostatistique de l'Institut Fédératif de Recherches en épidémiologie, sciences sociales et santé publique, Villejuif.

« Estimateurs *model-based* non paramétriques de la fonction de répartition d'une variable censurée à droite sur petits domaines », 2012, Séminaire de Statistique de TSE, Toulouse.

A.10 Activités d'enseignement

Formation initiale

Enseignement de la statistique du L1 au M2 à UT1 Capitole et en école d'ingénieur (INSA de Toulouse) : statistique descriptive et inférentielle, modèle linéaire, tests d'hypothèses, statistique mathématique, analyse des durées de vie, encadrement de projets et de stages.
Responsable pédagogique du magistère d'Economiste Statisticien depuis septembre 2017.

Formation continue

Cours bénévoles pour la SFdS sur les méthodes de base et sur les techniques avancées en analyse des données de survie, pour un public venant en majorité de l'industrie pharmaceutique.

B — Liste des publications

B.1 Publications dans des revues à comité de lecture

[1] **Leconte, E.**, Moreau, T et Lellouch, J. (1994), « The two-sample problem with multivariate censored data : a new rank test family », *Communications in Statistics : computation and simulation*, 23, 1061–1076, http://www.tandfonline.com/doi/abs/10.1080/03610919408813217.

[2] Aragon, Y., Haughton, D., Haughton, J., **Leconte, E.**, Malin, E., Ruiz-Gazen, A. et Thomas-Agnan, C. (2002), « Explaining the Pattern of Regional Unemployment : The Case of the Midi-Pyrénées Region », *Papers in Regional Science*, 82, 155–174, https://link.springer.com/article/10.1007/s101100200106.

[3] **Leconte, E.**, Poiraud-Casanova, S. et Thomas-Agnan, C. (2002), « Smooth Conditional Distribution Function and Quantiles under Random Censorship », *Lifetime Data Analysis*, 8, 229–246, https://www.ncbi.nlm.nih.gov/pubmed/12182120.

[4] Derzko, G. et **Leconte, E.** (2004), « Estimation non paramétrique d'incidences d'événements en compétition avec censure à droite », *Journal de la SFdS*, 145(1), 47–69, www.numdam.org/article/JSFS_2004__145_1_47_0.pdf.

[5] Derzko, G. et **Leconte, E.** (2004), « Estimation non paramétrique pour des événements répétés tronqués et censurés aléatoirement à droite – Applications biomédicales », *Journal de la SFdS*, 145(2), 77–101, www.numdam.org/article/JSFS_2004__145_2_79_0.pdf.

[6] Aragon, Y., Chambers, R., Casanova, S. et **Leconte, E.** (2005), « Conditional ordering using nonparametric expectiles », *Journal of Official Statistics*, 21(4), 617–633, ro.uow.edu.au\T1\guilsinglrighteis\T1\guilsinglrightpapers\T1\guilsinglright2516.

[7] Dupuy J.-F. et **Leconte, E.** (2009), « A study of regression calibration in a partially observed stratified Cox model », *Journal of Statistical Planning and Inference*, 139(2), 317–328, https://doi.org/10.1016/j.jspi.2008.04.024.

[8] Allera-Moreau, C., Rouquette, I., Lepage, B., Oumouhou, N., Walschaerts, M., **Leconte, E.**, Schilling, V., Gordien, K., Brouchet, L., Delisle, M.B., Mazieres, J., Hoffmann, S., Pasero, P. et Cazaux, C. (2012), « DNA replication stress response involving PLK1 CDC6, POLQ, RAD51 and CLASPIN upregulation prognoses the outcome of early/mid-stage non-small cell lung cancer patients », *Oncogenesis*, 1, e30, https://www.nature.com/articles/oncsis201229.

[9] Somda, S.M.A., **Leconte, E.**, Kramar, A., Penel, N., Chevreau, C., Delannes, M., Rios, M. et Filleron, T. (2014), « Determining the Length of Posttherapeutic Follow-up for Cancer Patients Using Competing Risks Modeling », *Medical Decision Making*, 34(2) : 168–179, http://journals.sagepub.com/doi/pdf/10.1177/0272989X13492015.

[10] Casanova, S. et **Leconte, E.** (2015), « A nonparametric model-based estimator for the cumulative distribution function of a right censored variable in a finite population », *Journal of Surveys : Statistics and Methodology*, 3, 317–338, https://academic.oup.com/jssam/article-abstract/3/3/317/1253765.

[11] Somda, S.M.A., **Leconte, E.**, Kramar, A., Boher, J.-M., Asselain, B. et Filleron, T. (2016), « Optimal scheduling of post-therapeutic follow-up of patients treated for cancer for early detection of relapses », *Statistical Methods in Medical Research*, 25(6) : 2457–2471, http://journals.sagepub.com/doi/abs/10.1177/0962280214524178?journalCode=smma.

[12] Somda, S.M.A., Florence, D., **Leconte E.** et Filleron T. (2016), « Oncologic surveillance after curative treatment : A probability-based approach should be preferred to a risk-based approach », *Urological Oncology : Seminars and Original Investigations*, 34(5), 244–245, https://www.ncbi.nlm.nih.gov/pubmed/27038697.

[13] Somda, S.M.A., Culine, S., Chevreau, C., Fizazi, K., **Leconte, E.**, Kramar, A. et al. (2017). « A statistical approach to determine the optimal duration of post-treatment follow-up : application to metastatic nonseminomatous germ cell tumors », *Clinical Genitourinary Cancer*, 15(2) : 230-236, http://dx.doi.org/10.1016/j.clgc.2016.07.023.

[14] Gilhodes, J., Zemmour, C., Ajana, S., Martinez, A., Delord, J.-P., **Leconte, E.**, Boher, J.-M. et Filleron, T. (2017), « Comparison of variable selection methods for high-dimensional survival data with competing events », *Computers in Biology and Medicine*, 91, 159–167, https://www.ncbi.nlm.nih.gov/pubmed/29078093.

B.2 Chapitres dans des ouvrages collectifs

[15] Dupuy, J.-F et **Leconte, E.** (2009), « Cox regression with missing values of a covariate having a non-proportional effect on hazard of failure », Mathematical Methods in Survival Analysis, Reliability and Quality of Life, Eds. Huber, C. et al., ISTE & Wiley.

B.3 Communications dans des conférences internationales avec comité de lecture et publications des actes

[16] Auriol, B., Garcia, F., Puech, L., Lagrange, S., Morer, C., Campardon, M., Rabat, O., Valentin, S., Perrin, O. et **Leconte, E.** (2004), « Agapè :

group telepathy. A long-term experimental series », *Proceedings of the 47th Annual Convention of the Parapsychological Association*, Vienne (Autriche), 325–346.

[17] Salameh, F., Picot, A., Chabert, M., **Leconte, E.**, Ruiz-Gazen, A. and Maussion, P. (2015), « Variable Importance Assessment in Lifespan Models of Insulation Materials : A Comparative Study », *Proceedings of 10th IEEE International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives*, Guarda (Portugal), 198–204, https://hal.archives-ouvertes.fr/hal-01375415.

B.4 Articles soumis ou en révision

[18] Cabarrou, B, Sfumato, P, **Leconte, E.**, Boher, J.-M., Filleron, T (2017), « ph2hetero : an R package for designing phase II clinical trials to target subgroup of interest in a heterogeneous population », Soumis à *Computer Methods and Programs in Biomedicine*.

B.5 Prépublications et travaux en cours

[19] Walschaerts, M., **Leconte, E.** et Besse, P. (2012), « Stable variable selection for right censored data : comparison of methods », https://arxiv.org/abs/1203.4928.

[20] Cabarrou, B., Dalenc, F., Somda, S.M.A., Kramar, A., Genre, L., Delord, J.-P., **Leconte, E.**, Boher, J.-M., Filleron, T. (2017), « Focus on an infrequently used quantity in the context of competing risks : the conditional probability function ».

[21] Somda, S.M.A., **Leconte, E.**, Solomiac, A., Gaston, C., Boher, J.-M., Filleron, T. (2017), « Confidence interval for survival functions : Comparison of different methods ».

[22] Casanova, S. et **Leconte, E.** (2017), « Nonparametric model-based estimators for the cumulative distribution function of a right censored variable in a small area ».

[23] Somda, S.M.A., Cabarrou, B., **Leconte, E.**, Boher, J.-M., Kramar, A., Filleron, T. (2017), « An algorithm to evaluate follow-up strategies after primary treatment in oncology by computer simulation ».