Mémoire présenté devant l'ENSAE ParisTech
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 14/02/2018

Par : **Tristan JUDD**

Titre : **Modélisation de la durée de maintien en arrêt de travail**

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière

M. Nicolas BARADEL

Entreprise : PRO BTP
Nom : Mme Béatrice DAY-PEDEUX
Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur du mémoire en entreprise :

Nom : Mme Béatrice DAY-PEDEUX
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Mots clés : incapacité, invalidité, apprentissage supervisé, modèle de taux, modèle de durée, censure, troncature, Kaplan-Meier, Cox, réseaux de neurones, mélange de lois

La modélisation de la durée des arrêts de travail est un enjeu central au bon provisionnement du risque. Les modèles usuels pour la durée des arrêts de travail sont basés sur l'évaluation des taux instantanés, historiquement afin de pouvoir prendre en compte les censures et troncatures. De nouvelles méthodes permettent néanmoins d'adapter les modèles centrés sur les durées à ces questions, ouvrant de nouvelles possibilités de modélisation en assurance de personnes. L'objectif de ce mémoire fut de se saisir de ces nouvelles opportunités pour proposer un modèle adapté aux spécificités de la législation française qui scinde les arrêts de travail en deux états. Un état d'incapacité pour un arrêt jugé court, d'un maximum de 3 ans, et un état d'invalidité pour des arrêts longs, s'étalant sur plusieurs années voire jusqu'à la retraite.

Une première partie du travail eut pour objectif de replacer les méthodes usuelles de construction de tables de taux, en particulier l'estimateur de Kaplan-Meier et le modèle de Cox, dans le cadre de l'apprentissage supervisé, permettant une comparaison des tables basée sur leur performance. Ce cadre permit aussi de proposer une méthode plus rigoureuse qu'une sélection visuelle pour le degré de lissage des taux obtenus par l'estimateur de Kaplan-Meier et de discuter de la sélection des variables dans le modèle de Cox.

Dans une seconde partie, nous avons implémenté et comparé par le biais de réseaux de neurones, outils/modèles peu envisagés encore en actuariat, deux méthodes d'adaptation des modèles à la censure et la troncature afin d'en cerner les avantages et inconvénients et de proposer une première implémentation concrète de tels modèles. Nous avons finalement proposé un modèle théorique basé sur un mélange de deux lois pour la durée des arrêts, classant les sinistres en terme de risque court et de risque long, et étudié comment celui-ci pourrait répondre aux questions que soulève la séparation en incapacité et invalidité. Comment évolue la répartition entre risque court et risque long des arrêts de travail lors du passage de l'incapacité à l'invalidité ? Quelle est la proportion restante de risque court/long lors de ce passage ? Si la durée en incapacité venait à être modifiée, comment modéliser une telle évolution ? Cependant, l'intérêt d'un tel modèle, en terme de performances, reste encore à être démontré.

Abstract

Keywords : disability, incapacity, invalidity, supervised learning, rate model, duration model, censoring, truncation, Kaplan-Meier, Cox, neural networks, mixture model

The modelization of duration is a central issue for the provisioning of work disability risk. Historically, rate models are used for this risk to take into account censored and truncated data. Recent methods allow us to adapt duration models to truncated and censored data, bringing new opportunities in life insurance. The aim of this work was to propose, through those recent methods, a model adapted to the specification of french legislation which splits the work stoppage into two states. A temporal disability state reflecting a short term risk called incapacity lasting at most 3 years, a permanent disability state called invalidity reflecting a long term risk.

In the first part of our work, we replaced the usual models in table construction, specifically the Kaplan-Meier estimator and Cox model, in a supervised learning framework, to allow a more rigorous comparison of the tables. We also proposed through this framework a more rigorous method for the degree of smoothing of rates obtained with the Kaplan-Meier estimator than the classic visual method and discussed variables selection in the Cox model.

In a second part, we implemented and compared through the use of neural networks, tools/-models still unusual in actuarial science, two different ways to adapt duration models to censored and truncated data, as a way to propose a first implementation and determine the pros and cons of those methods. Finally, we proposed a theorical study of a mixture model for duration, classifying claims in term of short risk and long risk, and how this model could answer questions concerning the separation in incapacity and invalidity. How does the repartition in short term risk and long term risk evolves through the change from incapacity to invalidity ? What is the remaining proportion of short/long term risk at the time of the transition ? If the maximum legal duration in incapacity came to change, how to modelize this evolution ? Yet, the impact of such model in term of performances still has to be demonstrated.

Table des matières

Remerciements

Je tiens à remercier ma responsable Béatrice Day-Pedeux pour avoir veillé au bon déroulement de mes travaux et pour son aide.

Je remercie l'équipe Prévoyance-Santé ainsi que l'équipe Transverse, en particulier tous ceux avec qui j'ai partagé le bureau ou les repas, pour les bons moments passés ensemble.

Je remercie enfin l'ensemble de la direction de l'actuariat pour l'ambiance de travail agréable

Introduction et problématique

Les provisions mathématiques des contrats traitant de l'arrêt de travail constituent une part conséquente du passif chez Pro BTP, s'élevant à plus de deux milliards d'euros. La détermination d'une table d'expérience la plus précise possible constitue alors un enjeu important, permettant de calculer la solvabilité au plus juste et d'optimiser les fonds propres.

L'essor de la data science ces dernières années ouvre de nouvelles possibilités. La construction de table, qui reposait sur quelques méthodes adaptées au problème, est désormais envisageable avec la majorité des méthodes d'apprentissage. La puissance de calcul disponible et les algorithmes mis au point permettent d'adapter facilement les différentes méthodes aux enjeux des données tronquées et censurées.

Ce mémoire s'inscrit dans une démarche visant à utiliser ces nouvelles méthodes afin d'améliorer la qualité de l'estimation des durées des arrêts de travail, la compréhension des arrêts et de la réglementation mise en place. Nous présenterons ici à la fois des modèles classiques de construction de table, replacés néanmoins dans un cadre plus général, ainsi que les outils d'utilisation des nouvelles méthodes disponibles et des modèles associés.

Après une présentation du domaine étudié, des données utilisées et des tables recherchées, nous décrirons dans une première partie le cadre statistique général et la construction de tables à l'aide de l'estimateur de Kaplan-Meier. Nous étudierons ensuite deux méthodes employées de lissage de tables, le lissage par splines cubiques et le lissage de Whittaker-Henderson, et enfin le modèle de Cox, méthode usuelle de prise en compte de variables supplémentaires pour compléter les tables.

Dans une seconde partie, nous présenterons les réseaux de neurones pour leur utilité en tant qu'outils dans la construction de modèles, ainsi que les modèles spécifiques qui leur sont associés. Nous présenterons ensuite une première implémentation de modèle de durée prenant en compte la censure et la troncature à travers une étude comparant deux méthodes différentes de gestion de la censure et de la troncature. Nous proposerons enfin un modèle de durée qui nous paraît spécifiquement adapté aux questions du maintien en arrêt en France, et comment celui-ci permettrait d'approfondir la compréhension de la séparation en état d'incapacité et d'invalidité.

1 Cadre de l'étude

1.1 Présentation de l'arrêt de travail

Notre travail fut effectué au sein de PRO BTP, un groupe paritaire à but non lucratif et professionel dans le domaine de la protection sociale du Bâtiment et des travaux publics. BTP-Prévoyance est une Institution de Prévoyance, assureur et gestionnaire des contrats conventionnels de prévoyance des ouvriers, ETAM et cadres, des contrats supplémentaires de prévoyance collective, des contrats de frais médicaux collectifs et qui proposent aussi des contrats de frais médicaux individuels pour les actifs et les retraités du secteur.

Malgré l'ouverture récente du secteur de la prévoyance à la concurrence, BTP-P en reste l'acteur majoritaire couvrant près de 1.6 million de salariés par an.

La prévoyance consiste à protéger les salariés des conséquences des aléas de la vie. Plus précisément, la loi Evin n^o89-1009 du 31 décembre 1989 définit que "La prévoyance regroupe les opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité ou des risques d'incapacité de travail ou d'invalidité ou du risque chômage ".

La majorité du portefeuille de BTP-P est composée des contrats collectifs, où une entreprise souscrit pour couvrir l'intégralité de ses salariés. Les garanties sont définies par des prestations financières dont l'objectif est la compensation, partielle, de la perte de salaire due à l'arrêt de travail qu'il soit temporaire ou permanent. Ces prestations sont versées d'une part par les branches maladies et accidents de travail de la Sécurité Sociale, d'autre part par les organismes complémentaires comme BTP-P.

De nombreux critères définissent les montants des remboursements. Nous n'en ferons pas ici la présentation précise, notre travail se focalisant sur la construction des tables plus que leur application pour le provisionnement et la tarification. En contrepartie de la couverture accordée, les salariés et les employeurs paient une prime. Ce montant dépend notamment de la durée des arrêts, de la cause (professionnelle ou vie privée) de l'arrêt et de la catégorie socioprofessionnelle de l'individu. Pour des contrats collectifs, le montant de cette prime ne peut pas varier selon des critères discriminatoires choisis par l'assureur. Rien n'empêche en revanche que la tarification proposée à chaque individu dépende de la composition démographique de l'entreprise, les informations sur l'âge et le sexe des bénéficiaires étant couramment employées.

Une distinction importante dans la législation française est effectuée entre une incapacité temporaire de travail et une incapacité permanente.

L'incapacité temporaire de travail (ou simplement incapacité par la suite) est l'état d'un assuré qui suite à une maladie ou un accident est contraint d'interrompre totalement ou partiellement son activité professionnelle. Les garanties étudiées consistent en un versement d'indemnités journalières pour un arrêt de travail avant l'âge de la retraite et de durée supérieure à la franchise du contrat. La durée maximale passée en incapacité fixée par la législation est de 3 ans. Au delà, par le biais

d'une procédure, l'arrêt est soit considéré comme terminé soit classé comme plus grave, risquant de s'étendre sur une échelle de temps bien plus éleveé.

L'état d'invalidité regroupe de tels problèmes, où l'assuré suite à une maladie ou un accident perd sa capacité de travail. La sécurité sociale classe distinctement en incapacité permanente de travail ou invalidité, selon que la maladie ou l'accident relève du caractère professionnel ou non. Les garanties étudiées consistent dans le versement d'une rente d'invalidité dont le montant dépend encore de la nature de l'arrêt, basé sur le calcul d'un taux d'incapacité pour les arrêts professionnels et sur un classement en catégorie pour les autres. Les catégories sont les suivantes :

— La catégorie 1 pour les personnes encore capables d'exercer une activité rémunérée

— La catégorie 2 pour les personnes invalides et incapables d'exercer une activité rémunérée.

— La catégorie 3 pour les personnes invalides et incapables d'exercer une activité rémunérée et d'accomplir la majorité des actes de la vie quotidienne nécessitant alors l'aide d'une tierce personne.

Le taux d'incapacité quant à lui est établi par le service médical de la Caisse d'Assurance Maladie et repose sur la la nature de l'infirmité, l'état général, l'âge, les facultés physiques et mentales de l'assuré ainsi que d'autres critères d'ordre professionnel.

Dans la majorité des cas, les arrêts en invalidité font suite à des arrêts en incapacité, puisque l'on préfère voir comment va évoluer l'état de l'individu plutôt que de désigner directement l'incapacité comme permanente.

Le regroupement des risques incapacité et invalidité forme le risque arrêt de travail. Dans notre étude, pour des raisons que nous détaillerons tout au long du rapport, nous nous intéressons à la durée totale des arrêts de travail en regroupant le temps passé en incapacité et invalidité des individus.

Majoritairement trois types de contrats sont proposés :

— Des contrats couvrant l'incapacité, avec une carence de 90 jours

— Des contrats couvrant l'invalidité, pour des causes non-professionnelles de catégorie 1, 2 et 3 ainsi que les causes professionnelles

— Des contrats où BTP-P s'engagent à verser à la place des entreprises les montants prévus par la loi de mensualisation du 19 janvier 1978. Concrètement, ces contrats correspondent à une couverture de l'incapacité (avec quelques jours de carence) sur des durées ne dépassant pas les 90 jours

Ces différents contrats permettent de couvrir et d'étudier le risque d'arrêt de travail dans sa globalité. Cependant, là où les deux premiers types de contrat concernent la grande majorité des entreprises du portefeuille, la population couverte par les contrats traitant de la mensualisation n'est pas la même. Seules des petites entreprises souscrivent à ce contrat, puisqu'elles sont plus intéressées par la couverture des arrêts de travail de moins de 91 jours que les grandes entreprises qui peuvent mutualiser elles-mêmes ce risque. Afin d'éviter un éventuel biais d'échantillonage, les arrêts concernant les contrats de mensualisation ne seront pas inclus dans l'étude.

La loi Evin a introduit les tables de maintien en incapacité et en invalidité pour le provisionnement. Elles furent établies par le BCAC (Bureau Commun des Assurances Collectives) en 1993 sur un portefeuille des grandes compagnies d'assurance française de l'époque, afin d'être représentatives des entreprises et catégories socioprofessionnelles couvertes. Ces tables correspondent au suivi de cohortes d'individus en incapacité et invalidité dans le temps. Elles admettent comme entrées l'âge au départ et le temps passé en incapacité/invalidité et indiquent le nombre d'individus restant en arrêt, les cohortes étant constituées à la base de 10 000 personnes. Ces tables permettent de lire directement, en divisant par 10 000, la fonction de survie des individus d'âge a à l'instant t, $S(t) = \mathcal{P}(X \geq t \mid \text{Age} = a)$, X étant le temps passé dans l'état de l'individu. Il est commun, comme nous le ferons dans cette étude, de travailler sur des tables représentant le taux de sortie $\mathcal{P}(X \leq t+1 \mid X > t \wedge \text{Age} = a)$ plutôt que la fonction de survie, le passage de l'une à l'autre s'effectuant facilement puisque $p_t^a = (S_t^a - S_{t+1}^a)/S_t^a$.

LOIS DE MAINTIEN EN INCAPACITE TEMPORAIRE (DEFINITION SECURITE SOCIALE)

ÂGE	MOIS 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
20	10 000	2 842	1 743	1 144	838	625	455	339	291	253	215	187	173	152	138	129	123	114	102	98
21	10 000	2 842	1 743	1 144	838	625	455	339	291	253	215	187	173	152	138	129	123	114	102	98
22	10 000	2 842	1 743	1 144	838	625	455	339	291	253	215	187	173	152	138	129	123	114	102	98
23	10 000	2 842	1 743	1 144	838	625	455	339	291	253	215	187	173	152	138	129	123	114	102	98
24	10 000	2 931	1 848	1 215	894	657	478	343	291	256	217	183	166	143	130	121	114	105	95	91
25	10 000	3 080	2 001	1 345	997	739	536	382	327	289	251	216	195	172	159	149	140	129	116	113
26	10 000	3 177	2 112	1 461	1 087	812	591	431	372	325	285	249	226	201	186	171	161	150	137	129
27	10 000	3 251	2 180	1 540	1 156	869	643	476	407	360	320	285	263	237	222	207	192	179	168	159
28	10 000	3 298	2 243	1 600	1 209	915	688	524	448	400	359	322	297	270	255	238	222	210	199	189
29	10 000	3 348	2 273	1 640	1 246	956	726	559	476	425	384	352	327	298	280	262	247	233	220	208
30	10 000	3 386	2 275	1 659	1 264	964	744	583	494	439	396	363	338	308	287	267	252	240	227	214
31	10 000	3 388	2 228	1 618	1 249	965	756	595	501	449	406	375	347	318	295	276	261	250	236	223
32	10 000	3 433	2 238	1 617	1 254	975	772	612	522	468	421	388	357	325	302	279	264	252	235	222
33	10 000	3 466	2 235	1 627	1 260	983	782	628	540	484	431	395	364	332	310	286	270	256	238	223
34	10 000	3 567	2 298	1 684	1 321	1 033	828	684	597	535	477	436	401	366	344	319	298	282	265	247
35	10 000	3 645	2 331	1 705	1 357	1 082	876	732	647	586	528	481	443	402	377	351	331	309	294	275
36	10 000	3 701	2 390	1 747	1 390	1 106	905	771	682	617	560	508	469	428	397	370	347	323	308	287
37	10 000	3 822	2 458	1 804	1 430	1 148	932	801	704	635	579	526	487	443	406	379	357	335	319	298
38	10 000	3 958	2 526	1 851	1 479	1 193	980	841	739	671	616	564	521	477	439	411	384	358	340	319
39	10 000	4 035	2 600	1 923	1 541	1 266	1 055	915	807	739	680	623	572	530	486	455	427	400	381	364
40	10 000	4 073	2 652	1 973	1 575	1 303	1 097	965	853	783	719	659	607	565	521	490	458	428	404	384

LOIS DE MAINTIEN EN INVALIDITE

ÂGE	AN 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
20	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
21	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
22	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
23	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
24	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
25	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
26	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
27	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
28	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
29	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
30	10 000	9 859	9 699	9 534	9 331	9 163	8 994	8 874	8 761	8 696	8 619	8 571	8 429	8 321	8 305	8 283	8 258	8 088	8 049	8 006
31	10 000	9 868	9 731	9 538	9 364	9 174	9 013	8 913	8 815	8 756	8 687	8 641	8 506	8 400	8 372	8 335	8 293	8 117	8 057	7 990
32	10 000	9 843	9 698	9 537	9 306	9 120	8 985	8 846	8 771	8 685	8 632	8 542	8 410	8 325	8 297	8 222	8 100	7 957	7 851	7 785
33	10 000	9 844	9 705	9 562	9 328	9 131	8 997	8 872	8 789	8 695	8 606	8 527	8 384	8 307	8 278	8 172	8 020	7 885	7 783	7 716
34	10 000	9 827	9 669	9 523	9 301	9 084	8 908	8 770	8 665	8 561	8 461	8 386	8 231	8 158	8 129	7 996	7 852	7 725	7 627	7 560
35	10 000	9 818	9 663	9 509	9 281	9 039	8 874	8 734	8 597	8 455	8 380	8 311	8 165	8 071	8 042	7 886	7 719	7 598	7 469	7 367
36	10 000	9 805	9 641	9 495	9 258	9 038	8 852	8 724	8 573	8 456	8 306	8 222	8 067	7 978	7 948	7 778	7 622	7 538	7 415	7 310
37	10 000	9 801	9 640	9 501	9 269	9 051	8 861	8 757	8 601	8 445	8 280	8 154	7 995	7 893	7 864	7 708	7 514	7 351	7 225	7 116
38	10 000	9 787	9 620	9 462	9 253	9 050	8 864	8 761	8 590	8 416	8 254	8 141	7 982	7 890	7 843	7 679	7 479	7 317	7 164	7 052
39	10 000	9 751	9 566	9 414	9 214	9 018	8 861	8 762	8 586	8 397	8 218	8 101	7 936	7 857	7 805	7 631	7 414	7 238	7 072	6 952
40	10 000	9 751	9 562	9 424	9 214	9 012	8 843	8 730	8 537	8 359	8 201	8 085	7 881	7 778	7 694	7 515	7 279	7 105	6 940	6 819

FIGURE 1: *Extrait des tables du BCAC pour le maintien en incapacité et invalidité, source (Bagui, 2013)*

Les tables construites sur toute la population française ne sont pas suffisamment représentatives du secteur du BTP. La répartition en catégorie socioprofessionnelle est différente et le risque couvert n'est globalement pas de la même nature, avec les accidents de travail sur les chantiers par exemple. Il est recommandé alors d'utiliser les informations des sinistres déjà observés pour construire des tables de maintien, appelées tables d'expériences, afin que celles-ci tiennent compte des spécificités du portefeuille de BTP-P.

1.2 Traitement des données

Les données nécessaires à la création des tables furent extraites en essayant de respecter le plus fidèlement possible les recommandations de l'Institut des Actuaires (Aubin and Rolland, 2010). Les logiciels utilisés furent le logiciel SAS, outil dont l'utilisation est répandue au sein de l'entreprise et directement connecté aux entrepôts de données, et l'outil de système de gestion, permettant d'afficher les informations détaillées pour tout l'historique d'un adhérent et des sinistres, servant de référence pour les vérifications et la compréhension des anomalies.

Pour donner un exemple de l'importance des différentes vérifications à effectuer sur les données, nous traiterons des rechutes de sinistre. Les rechutes sont codées par l'ajout d'un arrêt dans la base, néanmoins afin de faciliter des traitements qui ne sont pas en lien avec ces travaux, la date de fin de rechute et la date de fin du sinistre d'origine sont modifiées. La vérification qu'un même individu n'avait pas plusieurs arrêts pour incapacité au même moment nous a permis de détecter cette spécificité et de la corriger.

De nombreuses bases de données prétraitées étaient à notre disposition, regroupant les informations disponibles à différentes échelles. En particulier, deux tables recensent l'ensemble des informations collectées sur 10 ans pour les sinistres incapacité et invalidité qui sont mises à jour mensuellement. Afin de collecter une quantité suffisante d'informations pour la construction des tables, en supposant que les variations de comportement sur la période considérée sont suffisamment faibles, nous avons sélectionné une période d'observation s'étendant de début 2008 à fin 2015 pour l'incapacité et de 2007 à 2015 pour l'invalidité.

Ces tables comprennent la date de fait générateur de sinistre, la date de fin de sinistre, de début et de fin d'indemnisation et un indicateur pour rattacher le sinistre à l'assuré concerné. Cet indicateur nous a permis de récolter les informations nécessaires sur l'individu, en particulier la date de naissance, de départ à la retraite et de décès (si elles existent) afin de corriger la date de sortie au besoin et de déterminer la cause de sortie (fin d'arrêt ou censure). L'indicateur a aussi rendu possible la jonction avec la table des droits (droits gratuits compris) afin d'éliminer les sinistres commençant hors périodes de droits.

Parmi les autres informations récupérées nous comptons la cause (privée ou professionnelle) du sinistre, le sexe de l'individu et sa catégorie socioprofessionnelle. De nombreuses informations supplémentaires, qui pour la plupart ne seront pas utilisées par la suite par manque de temps, furent aussi ajoutées. Une base de données sur les entreprises fut utilisée pour calculer le département du sinistre ainsi que la taille (petite ou grande) de l'entreprise au moment du sinistre. Nous avons aussi récupéré de nombreuses informations sur le site de l'INSEE, nous focalisant sur des indicateurs socio-économiques géographiques et des indicateurs trimestriels couvrant les secteurs du Bâtiment et des Travaux publics.

Les tables sont normalement construites sur la durée passée dans l'état, avec une table pour l'incapacité et une table pour l'invalidité. Bien que pour la présentation et les besoins des différentes méthodes d'apprentissage nous effectuons une séparation similaire en terme de durée, nous avons

choisi de travailler sur la durée totale de l'arrêt, regroupant le temps passé en incapacité et en invalidité.

Il faut rappeler que conformément à la loi Evin et à des décisions de justice plus récentes [1], pour les contrats couvrant l'incapacité et l'invalidité (représentant la grande majorité du portefeuille) les assureurs se doivent de rembourser les incapacités prolongées en invalidité même après la résiliation ou le non renouvellement du contrat , la durée totale de l'arrêt apparaissant dans le provisionnement.

Le choix de travailler sur la durée totale paraît néanmoins à première vue contre productif puisque la tarification et le maintien dépendent du temps passé en incapacité et invalidité, séparément. Il est possible de compléter la table de durée totale d'arrêt avec une table de passage d'incapacité en invalidité, qui serait assez similaire à la table de passage utilisée actuellement pour compléter les tables de maintien. Finalement, les deux représentations, durée totale d'arrêt avec une table de passage ou durée en incapacité, durée en invalidité avec une table de passage, sont assez équivalentes.

Nous pensons que la construction d'une seule table repose sur l'étude d'une grandeur plus "naturelle" qu'est la durée d'arrêt, là où les tables d'incapacité et d'invalidité s'intéressent à des états définis par la législation. Cette séparation en deux états des arrêts découle d'une représentation en deux types des risques encourus par les individus : un risque court évoluant en mois et un risque long qui s'étend sur plusieurs années. La loi observée dans chaque état est dans cette représentation un mélange de lois tronquées des deux types : l'incapacité va comprendre les risques courts ainsi que la majorité des départs des risques longs alors que l'invalidité contient la suite des risques longs et la fin de quelques risques courts (ceux-ci étant malgré les vérifications classés en invalidité).

Nous présentons dans la partie II 7.1 de ce mémoire un modèle de mélange de deux lois visant à reproduire cette distinction entre risques longs et risques courts afin d'améliorer les performances et d'étudier la limite législative de 3 ans en incapacité.

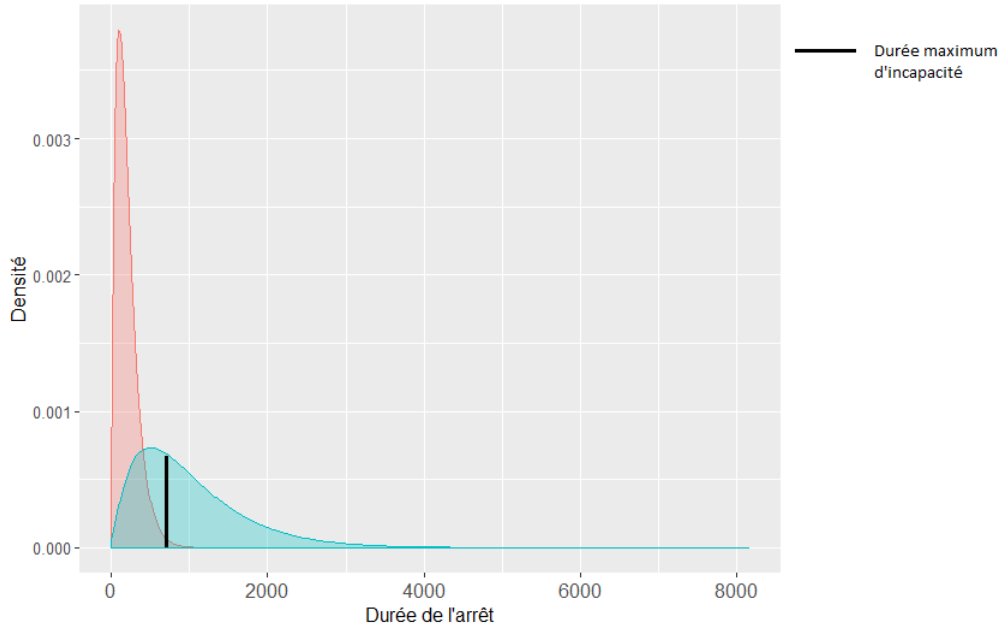


FIGURE 2: *Illustration graphique de la répartition en risque court, risque long et de la délimitation imposée par la législation*

1. Arrêt n° 222 du 16 janvier 2007, Cour de cassation

Ce traitement joint limite malheureusement la portée de l'étude, les tables d'expérience construites n'étant pas directement comparables avec les tables du BCAC car ne s'intéressant pas au même phénomène. Leur comparaison nécessiterait de les compléter avec une table de passage pour comparer les provisionnements et tarifications, ce qui soulève encore d'autres questions (voir 2.1).

Les tables furent finalement extraites afin de réaliser la suite de l'étude avec le logiciel R.

1.3 Statistiques descriptives

Les données utilisées dans cette étude sont constituées des informations sur la date de début et de fin d'observation, la cause de fin d'observation (fin d'arrêt ou censure), l'âge à la date de fait générateur, la cause de l'arrêt ainsi que le sexe et la catégorie socioprofessionnelle de l'individu. La partie incapacité est composée de 199470 arrêts et la partie invalidité de 33011, on s'attend alors à une plus grande instabilité des estimations sur la partie invalidité.

Variable	Modalité	Moyenne	Variance	Censure (%)
Sexe	Homme	326	579	26.9
Sexe	Femme	291	517	27.2
Cause d'arrêt	Privée	342	680	31.2
Cause d'arrêt	Professionnelle	285	233	16.0
Catégorie socioprofessionnelle	Cadre	305	493	33.4
Catégorie socioprofessionnelle	Etam	297	514	29.5
Catégorie socioprofessionnelle	Ouvrier	328	586	26.2

Table 1: *Informations sur les variables explicatives supplémentaires : moyenne, variance (hors censure) du jour de fin d'observation et pourcentage d'observations censurées*

À première vue en ne regardant que la moyenne du jour de fin d'observation, pouvant varier jusqu'à une cinquantaine de jours entre les modalités, nous serions tenté de dire que ces variable supplémentaires contiennent des informations influençant la fin de l'arrêt. La variance et le pourcentage de censure élevés, non pris en compte, nous rappellent que l'on ne peut pas tirer de conclusion quant à l'impact de ces variables sans outils statistiques adaptés.

Pour la suite de l'étude, nous avons scindé les données en deux parties afin d'en faciliter la compréhension et permettre la bonne utilisation des méthodes d'apprentissage. Une première partie mesurée en mois jusqu'au 1096^{eme} jour d'observation, assimilable à de l'incapacité, et une autre partie en année couvrant la suite des sinistres, assimilable à l'invalidité.

FIGURE 3: *Répartition des individus sur la partie incapacité (gauche) et invalidité droite, concernant le sexe (haut), la catégorie socioprofessionelle (milieu) et la cause de l'arrêt (bas)*

Les proportions d'hommes et d'ouvriers du portefeuille sont plus élevées que la répartition en population générale et sont relativement stables entre les deux parties. Il est aussi intéressant de noter que la proportion d'arrêt de cause professionnelle est environ 5 fois moins importante en invalidité qu'en incapacité : on en déduit que les taux de passage d'incapacité en invalidité, et donc la durée des arrêts, sont distincts entre les deux causes d'arrêts.

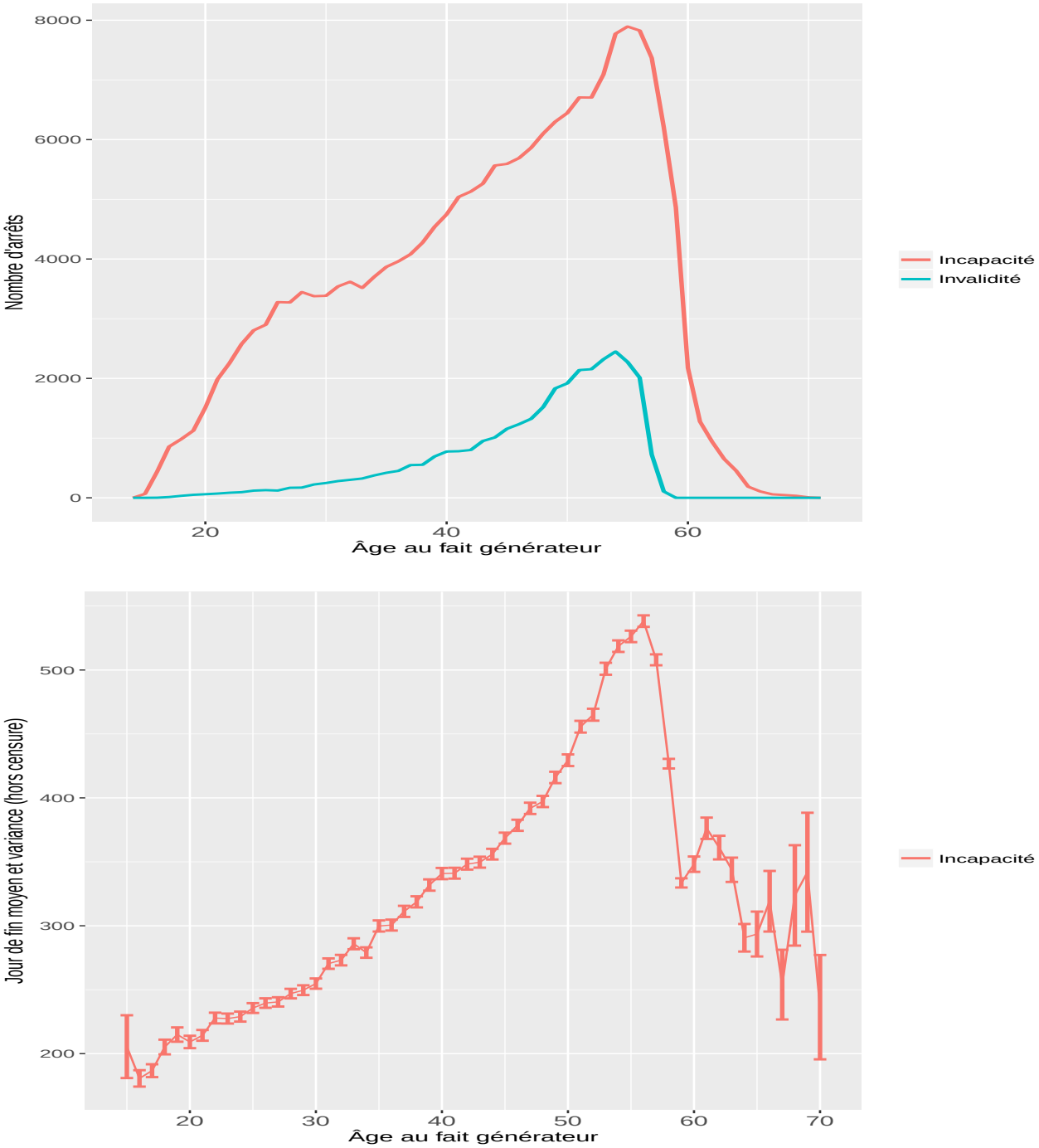


FIGURE 4: *Répartition des sinistres par âge pour la partie incapacité et invalidité (haut) et jour moyen et variance (hors censure) de fin d'observation sur l'incapacité*

La majorité du portefeuille se concentre sur les individus âgés de 40 à 55 ans, avec une répartition plus uniforme pour l'incapacité : les individus de 60 ans sont environ deux fois plus nombreux que ceux de 30 ans, là où la proportion atteint les 5 fois plus nombreux pour l'invalidité.

L'étude du jour de fin moyen en fonction de l'âge est fortement à nuancer, étant donné que la censure est corrélée à l'âge. Concernant l'incapacité, outre la variance importante dûe au faible nombre d'individus aux bouts de courbes (à 16 ans et pour les âges de plus de 62 ans, post date légale de départ à la retraite), nous observons une croissance du jour de fin moyen jusqu'à 55 ans puis sa décroissance. Encore une fois les censures n'étant pas prises en compte, ces représentations graphiques ne sauraient faire l'objet d'une véritable interprétation.

Première partie

Apprentissage supervisé et modèles de taux

2 Cadre statistique et estimation des taux de sortie

2.1 Principaux enjeux des méthodes d'apprentissage supervisé

2.1.1 Décomposition biais-variance

L'objectif en apprentissage supervisé est d'arriver à prédire le mieux possible une variable aléatoire Y à partir d'autres variables $X = (X_1, ..., X_p)$. Dans notre cas, nous cherchons à déterminer les taux de sortie à partir des caractéristiques disponibles des individus et les différentes périodes de temps considérées. Mathématiquement, le problème se pose comme la recherche d'une fonction qui aux données renverrait une prédiction qui minimiserait un certains critère $\mathcal{C}$ (qu'on nommera par la suite aussi erreur) : $f \in \arg\min\ \mathcal{C}\ (Y, f(X))$.

Concrètement, nous ne sommes pas capable de calculer cette fonction exactement. Nous disposons uniquement de n observations (les différents individus) supposées correspondre à des réalisations de variables aléatoires indépendantes et de même loi que X, et remplaçons l'erreur par son estimation empirique. L'approximation de la fonction se fait en deux temps :

1. La minimisation s'effectuant sur un nombre fini d'éléments, il y a une infinité de fonctions qui minimise ce critère empirique. Nous restreignons l'espace de recherche à un sous-ensemble choisi, qui variera selon les méthodes envisagées : $f(x) \approx \hat{f}(x, \theta)$ où $\theta \in \Theta$ ensemble des paramètres possibles de l'espace choisi.
2. La fonction choisie de ce sous-espace est celle minimisant ce critère empirique, il s'agit donc d'une estimation de la fonction de ce sous-espace minimisant l'erreur théorique : $\hat{f}(x, \theta) \approx \hat{f}(x, \hat{\theta})$.

Cette séparation en deux temps se répercute en termes distincts sur l'erreur, on parle alors de décomposition biais-variance. En pratique, il est difficile d'arriver à déterminer un sous-espace parfaitement adapté au problème, le statisticien devant effectuer un arbitrage :

— Un sous-ensemble restreint entraînera une stabilité de l'estimation de la fonction et donc une variance faible, mais des hypothèses erronées peuvent causer un biais important. La méthode est alors incapable d'utiliser correctement l'information contenue dans les données : on parle de sous-apprentissage

— Au contraire un sous-ensemble très vaste permettrait d'obtenir un biais presque nul. Cependant l'absence d'hypothèses clées peut amener le modèle à ne pas distinguer l'information pertinente du bruit des données, causant une variance élevée : on parle de sur-apprentissage

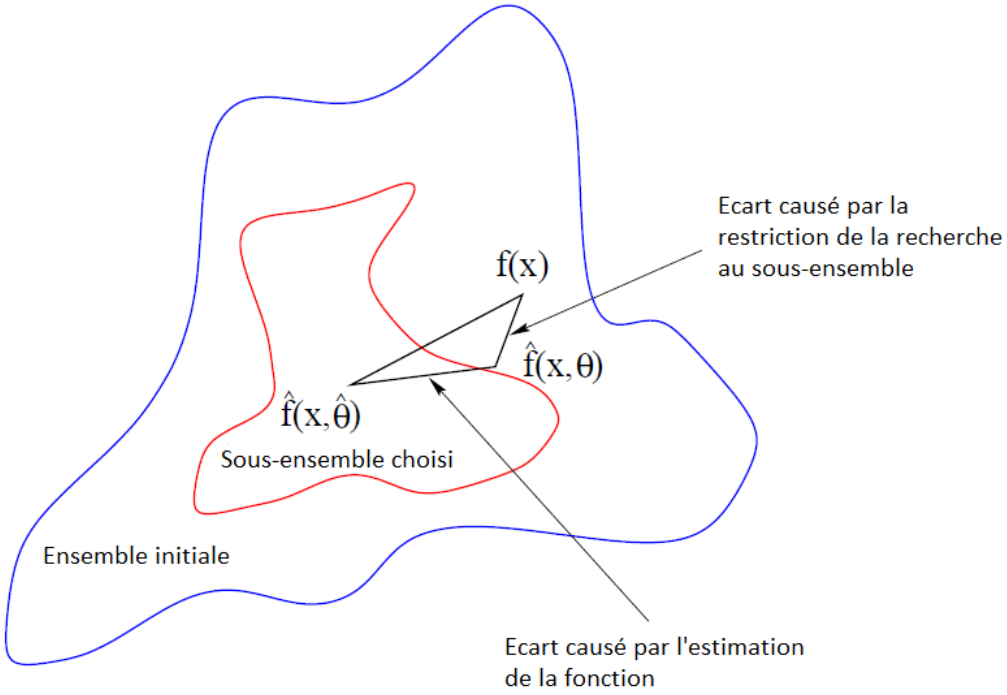


FIGURE 5: *Explication graphique des écarts vis à vis de la solution théorique idéale. Source : (de Freitas, 2000)*

2.1.2 Mesure des performances et calibrage des paramètres

Usuellement, la construction d'une table de maintien ne passe pas par ce cadre d'apprentissage supervisé et la définition d'un critère. La qualité d'une table d'expérience est déterminée indirectement par son impact sur le provisionnement et la tarification.

Cependant, la mesure par l'impact sur le provisionnement et la tarification est possible car le nombre de variables utilisées pour la construction des tables est relativement petit : pour des tables construites en grande dimension, l'impact sur chaque sous-segment possible n'est pas vérifiable, et l'étude à l'échelle agrégée imprécise puisque la compensation des erreurs est importante en assurance mais n'apparaît pas dans la définition d'un critère statistique viable (annexe 8.A). De plus, la grande majorité des méthodes d'apprentissage, non utilisées actuellement pour la construction de table, repose sur la définition d'un tel cadre. Nous espérons que l'usage de ces méthodes, couplées à l'introduction d'une quantité importante d'informations, améliore la qualité des tables construites ainsi que notre compréhension de l'arrêt de travail. À noter cependant l'existence de mesure spécifiquement adaptée à la tarification et au provisionnement (annexe 8.B), utilisée souvent en IARD.

Un problème plus profond avec l'évaluation des tables par ce genre de mesure, est que les tables en elles-mêmes doivent être complétées (par une table de passage d'incapacité en invalidité notamment) pour être évaluées, si bien qu'il est difficile de distinguer dans la mesure des performances les qualités provenant de la table de maintien, de la table de passage ou de la bonne combinaison des deux.

Nous dirons qu'une méthode a de meilleures performances qu'une autre si sa fonction estimée aboutit à une erreur empirique plus faible que celle de l'autre méthode. Pour ce faire, il est courant de séparer aléatoirement les données en échantillon (base) d'apprentissage sur lequel la fonction est

estimée ($\hat{f}(X,\hat{\theta}) = \hat{\mathrm{E}}[\hat{f}(X,\theta) \mid \mathcal{A} = \mathrm{A}]$) et en échantillon (base) test où l'erreur est approximée ($\hat{\mathcal{C}}(Y,\hat{f}(X))$, qui se réécrit bien souvent $\hat{\mathrm{E}}[\mathcal{L}(Y,\hat{f}(X))]$ où $\mathcal{L}$ fonction de perte).

En effet, l'usage du même échantillon pour l'estimation de la fonction et l'estimation de l'erreur aboutit généralement à une sous-estimation de cette erreur, les modèles effectuant du sur apprentissage étant favorisés. L'exemple ci-dessous illustre ce phénomène, l'échantillon correspondant à des réalisations indépendantes et identiquement distribués de $Y \sim \mathcal{U}\{-1;1\}$ et $X \sim \mathcal{U}\{0;...;5\}$ indépendantes. Le modèle idéal étant $\mathrm{E}[Y \mid X] = \mathrm{E}[Y] = 0$, nous présentons un modèle où apparaît un sur apprentissage, et dont l'erreur est nulle si les échantillons d'estimation de la fonction et de l'erreur sont les mêmes.

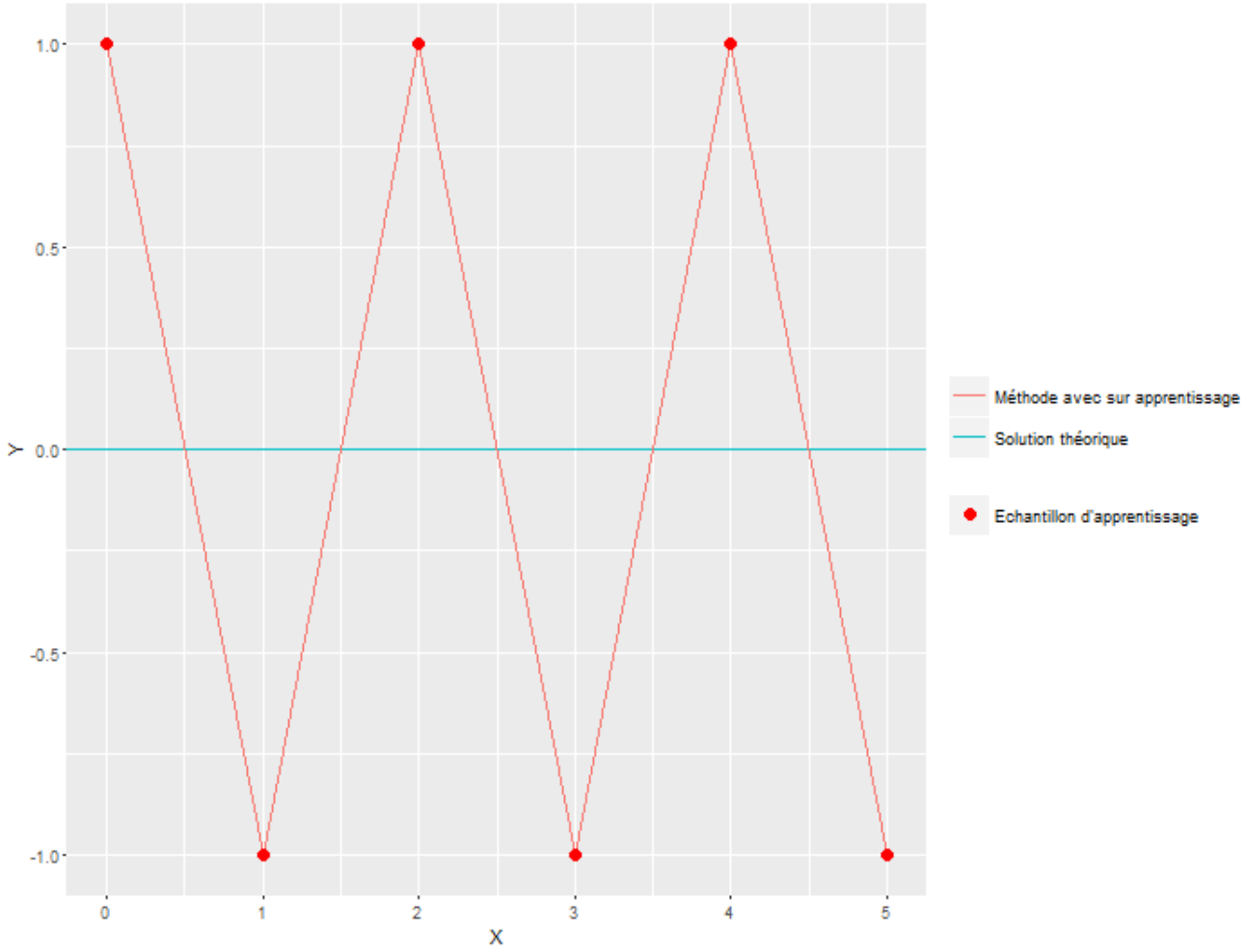


FIGURE 6: *Exemple de solution d'erreur nulle sur l'échantillon d'apprentissage comparée à la solution théorique optimum*

Beaucoup de méthodes d'apprentissage reposent aussi sur des paramètres, nommés hyperparamètres, qui ne varient pas au cours de l'estimation de la fonction et sont à déterminer en amont. Il s'agit alors d'une estimation supplémentaire, portant cette fois-ci sur l'espace des fonctions. Concernant la construction de table, il est courant d'utiliser des tests afin de déterminer ces paramètres, tels que les coefficients de pénalisation de lissage ou les variables à inclure au modèle de Cox. Dans cette étude, nous estimerons les paramètres par une séparation en base de test et d'apprentissage, de telle sorte que les paramètres sélectionnés minimisent l'erreur. Nous reviendrons sur les différences entre

cette méthode et l'estimation par des tests aux différentes occasions. S'agissant d'une estimation, les données pour déterminer ces paramètres ne doivent pas être utilisées pour l'estimation de l'erreur.

Il est commun (Hastie et al., 2008) de procéder à une segmentation en trois parties : base d'apprentissage pour l'estimation de la fonction, base de validation pour l'estimation de l'erreur pour les différents hyperparamètres possibles et base de test pour l'estimation de l'erreur de la méthode pour les hyperparamètres sélectionnés.

La méthode utilisée pour comparer les différentes tables de maintien est la suivante :

répartition aléatoire des individus en base d'apprentissage (50%),
de validation (25%) et base de test (25%)
pour chaque *méthode d'apprentissage* **faire**
  pour chaque *combinaison d'hyperparamètre envisagée* **faire**
    estimation de la fonction sur l'échantillon d'apprentissage
    estimation de l'erreur sur l'échantillon de validation
  fin
  sélection des hyperparamètres minimisant l'erreur sur l'échantillon de validation
  estimation de la fonction sur la base d'apprentissage pour les hyperparamètres choisis
  estimation de l'erreur sur la base de test
fin
comparaison des méthodes à partir des erreurs sur la base de test

Les paramètres variant d'un modèle à l'autre, ceux-ci seront comparés pour des temps de calcul équivalents.

Il existe cependant d'autres méthodes qu'une séparation en base d'apprentissage et base de test pour estimer l'erreur : la validation croisée à 5 échantillons ressort pour l'estimation des hyperparamètres sur certains exemples (pour la méthode de gradient boosting dans (Ridgeway, 2007)).

Une meilleure approximation de l'erreur qui ne dépendrait plus de l'échantillon d'apprentissage serait $\mathrm{E}_{\mathcal{T}}\left[\mathrm{E}_{X,Y}[\mathcal{L}(Y, f(X)) \mid \mathcal{T}]\right]$ où $\mathcal{T}$ est un échantillon d'apprentissage.

La validation croisée consiste à s'en rapprocher en formant des jeux de données distincts en séparant notre base en 5 échantillons : 4 d'entre eux sont utilisés comme base d'apprentissage et le dernier en base de test. Ce procédé peut être répété cinq fois pour avoir cinq bases d'apprentissage différentes. L'espérance précédente est ensuite approximée par la moyenne des erreurs sur les bases de test. Les échantillons ainsi formés ne constituent cependant pas de "véritables" échantillons, dans le sens où ceux-ci ne sont plus indépendants, la méthode étant à rapprocher des estimations de type bootstrap.

Hors du cadre de l'étude, qui nécessite l'usage d'une base de test et des temps de calcul rapides pour comparer les méthodes, nous conseillons d'utiliser une validation croisée pour la détermination des hyperparamètres puis de se servir de l'ensemble des données disponibles pour construire la table de maintien.

2.2 Cadre pour l'estimation des taux

2.2.1 Mesure de l'erreur par maximum de vraisemblance

Afin de simplifier les notations, nous omettrons les informations complémentaires disponibles X sur l'individu, la variable Y considérée ci-dessous correspondant plutôt à $\mathrm{E}[Y \mid X]$.

Soit $(t_1, ..., t_u)$ l'ensemble des dates de départ (et de fin) des périodes pour lesquels les taux sont calculés. Le taux de sortie p_{t_i} à la période $[t_i; t_{i+1}]$ n'est rien d'autre que $\mathcal{P}(Y \leq t_{i+1} \mid Y > t_i)$ où Y est la date de fin d'arrêt de l'individu. À chaque période t_i, l'individu se voit donc dans la possibilité de sortir avec la probabilité p_{t_i} ou de rester en arrêt avec la probabilité $1 - p_{t_i}$. En utilisant la définition de la probabilité conditionnelle, la probabilité pour un individu présent en t_j de finir à la période t_i est $\mathcal{P}(Y \leq t_i \mid Y > t_j) = \prod_{k=j}^{i-1}(1 - p_{t_k})p_{t_i}$. La vraisemblance pour un tel individu est alors équivalente à celle de k tirages indépendants de loi binomiales de paramètres $p_{t_k}, ..., p_{t_i}$. En supposant les durées observées des différents individus indépendantes, en notant $Y_{t_i}^k$ la variable aléatoire indiquant si l'individu k est sorti ou non au cours de la période $[t_i; t_{i+1}]$ et $\mathbb{1}_{t_i}^k$ la variable indiquant si l'individu k était présent ou non sur la période, la log-vraisemblance pour n individu peut s'écrire :

$$\sum_{k=1}^{n} \sum_{i=1}^{u-1} \mathbb{1}_{ti}^k \left[Y_{t_i}^k \log(p_{t_i}) + \left(1 - Y_{t_i}^k\right) \log\left(1 - p_{t_i}\right) \right]$$

Nous avons alors choisi d'utiliser comme critère l'opposé de cette log-vraisemblance (de sorte à avoir un critère à minimiser) pour la comparaison des méthodes et le calibrage des hyperparamètres.

L'évaluation des prédictions sur une base de test étant peu commune dans la construction de table de taux, nous avons dû implémenter les fonctions correspondantes sous R. Nous avons utilisé des fonctions de conversion, afin de transformer les tables de taux obtenues, disponibles dans des objets complexes difficilement manipulables et contenant d'autres informations superflues, au format de table de données (dataframe).

Notre première fonction pour calculer l'opposé de la log-vraisemblance à partir des observations et d'une table de taux fut un échec. Celle-ci appliquait à chaque ligne de la table des sinistres une fonction, qui à partir de la date de début et de fin, calculait un dataframe dupliquant la ligne de l'individu pour chaque période concernée, puis calculait la log vraisemblance pour cet individu. Le résultat était alors beaucoup trop long, l'évaluation de la log vraisemblance d'un modèle étant plusieurs ordres de grandeur au dessus de la construction du modèle.

Ce sont les différentes transformations réalisées qui dans cet essai prenaient du temps. La place en mémoire vive étant suffisante, nous avons pu réduire drastiquement la vitesse de calcul à l'aide du package *data.table* (Dowle and Srinivasan, 2017), conçu pour accélérer la gestion des bases de données, en construisant une unique table recensant l'ensemble des périodes plutôt que de construire une table par arrêt, réduisant les appels de fonction et le nombre de construction et destruction d'objets en mémoire. Ensuite, plutôt que de faire correspondre un taux à chaque individu, il nous est apparu utile de regrouper pour les mêmes périodes et les mêmes caractéristiques individuelles ces données, la contribution de n individus de même caractéristique à la log vraisemblance étant n fois

celle d'un individu ayant ces caractéristiques. Finalement, la base des individus agrégée fut stockée plutôt que d'être recalculée à chaque évaluation de table.

Cette implémentation n'est possible que lorsque la taille du jeu de donnée et le nombre de variables explicatives sont suffisamment petits pour que les bases réalisées (tables de taux et calcul intermédiaire) tiennent en mémoire. Même si certains packages existent pour la gestion de base de données ne tenant pas en mémoire, comme le package *ff* (Adler et al., 2014) fonctionnant avec un jeu d'écriture sur le disque et de mise en mémoire similaire à SAS, les temps de calcul ne sont pas comparables et auraient nécessité une toute autre organisation.

2.2.2 Censure et troncature

La particularité historique la plus importante séparant les techniques usuellement employées pour la construction des tables de maintien aux autres techniques de machine learning, employées en IARD par exemple, est que la construction des tables appartient au domaine plus vaste des phénomènes de survie : ceux-ci nécessitent des techniques adaptées à la gestion des censures et troncatures. Nous nous limiterons ici à une brève présentation des types de censure et troncature rencontrées dans nos données qui sont la censure à droite de type I et la troncature à gauche.

La censure à droite consiste en une information partielle sur la durée de l'arrêt : certains arrêts ne sont pas observés jusqu'au bout, cependant nous avons quand même une date de fin d'observation et pouvons en conclure que la date de fin d'arrêt est plus grande que celle-ci. L'observation de la date de fin d'arrêt Y est alors remplacée par l'observation de $\min(Y, C)$ où C date de censure et de $\delta = \mathbb{1}_{Y \leq C}$, indiquant si la donnée est censurée ou non.

Dans nos données, nous observons deux causes principales de censures : la censure causée par l'observation sur un intervalle de temps, tout arrêt continuant au delà de fin 2015 étant censuré, et la censure pour départ à la retraite. Le choix de considérer les départs à la retraite comme une observation censurée plutôt que comme une sortie d'arrêt est discutable, cette seconde option permettrait notamment une construction de table plus facile et moins sujette aux erreurs (sortie non reconnue pour départ à la retraite) mais présente le désavantage d'être plus sensible à la législation concernant le départ à la retraite. Le traitement du départ à la retraite comme cause de censure nécessite de compléter la table construite (par une table de départ à la retraite ou un âge de départ fixé) afin de tarifier et de provisionner.

Une dernière cause traitée en tant que censure dans cette étude, et habituellement non considérée comme telle, est la sortie au bout des 3 ans d'incapacité. Traiter ces cas comme des fins d'arrêts créerait un pic de taux qui perturberait les méthodes de lissage et serait contraire aux objectifs des travaux menés dans la partie II, qui ont pour objectif de retrouver ces sorties correspondant à des arrêts suffisamment longs pour durer plus de 3 ans, mais pas suffisamment graves pour passer en invalidité. Tout dépend une fois de plus du phénomène que l'on cherche à étudier, et ce choix nécessite un traitement supplémentaire afin de tarifier et provisionner.

La troncature à gauche est une information manquante sur la durée de l'arrêt : plutôt que d'étudier la loi sur la durée des arrêts directement, nous étudions la loi sachant qu'une partie des données a

été tronquée. Dans notre base de données, nous observons deux causes de troncatures à gauche. La première est liée à l'observation des données sur un intervalle de temps, les arrêts étant tronqués à la date de début d'observation et les arrêts s'étant terminés avant n'étant pas présents. La deuxième provient de l'usage d'une carence dans les contrats : les remboursements n'ayant lieu majoritairement qu'à partir de 90 jours, les arrêts plus courts ne sont pas tous déclarés à l'organisme. Cette carence de 90 jours est cumulable sur l'année pour certains contrats, et l'on dispose dans notre base de certains arrêts courts, pour les individus ayant faits des arrêts multiples. Afin de ne pas complexifier la loi observée en présence de troncature, nous avons choisi d'imposer une troncature fixe de 90 jours, et avons éliminé ces contrats de la base.

La censure peut être vue comme une information partielle du phénomène qui nous intéresse, alors que la troncature est un biais d'échantillonage qui constitue une information manquante, une contrainte conditionnant la loi observée. La loi observée en présence de troncature est une loi conditionnelle à la troncature des données, là où un arrêt censuré fournit une information sur la loi recherchée, mais une information seulement partielle.

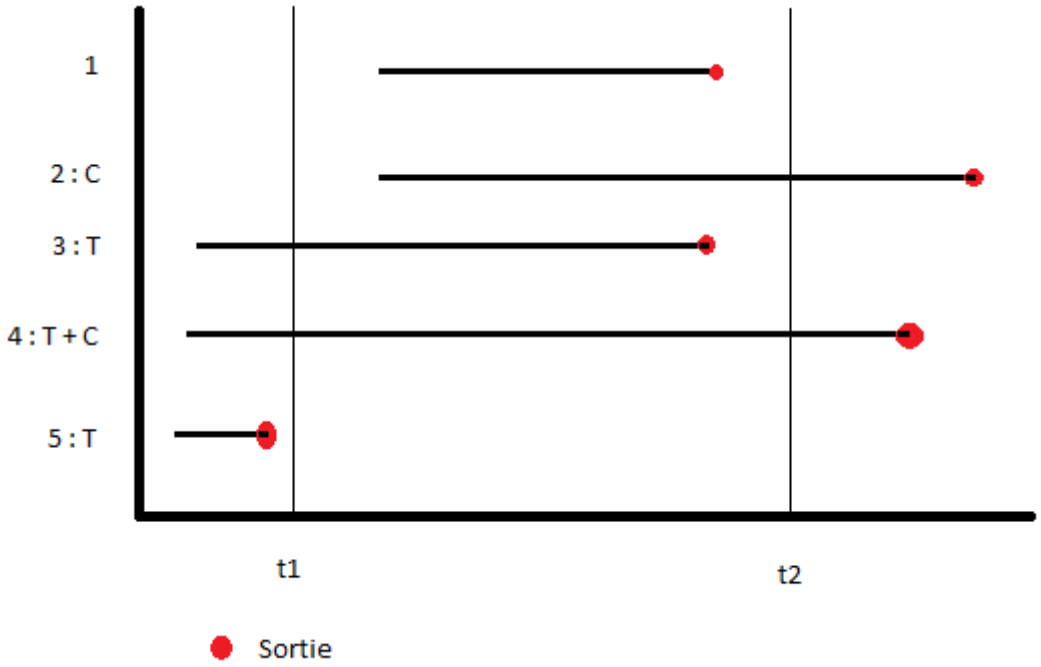


Figure 7: *Les différentes possibilités dans l'observation de données sur un intervalle : observation totale (1), censurée (2), tronquée (3), censurée et tronquée (4) et totalement tronquée donc non observée (5)*

La condition de présence au cours de la période utilisée dans notre formule de log vraisemblance est dans ce cas très restrictive : au cours d'une période, uniquement les individus présents au départ de la période et présents aussi à la fin de la période ou sortant pour une fin d'arrêt sont pris en comptes. Les individus commençant à être observés au cours de la période (troncature à gauche) ou sortant pendant la période pour une cause différente que la fin de l'arrêt (censure à droite), ne sont pas pris en compte dans la vraisemblance. Cette condition est très discutable et restrictive, tant la présence de censures et troncatures est importante dans la base et nous prive de l'utilisation d'une partie de l'information disponible, mais est nécessaire si l'on ne souhaite pas utiliser d'hypothèses supplémentaires.

La construction d'une table de taux étant l'étude discrétisée d'un phénomène continu, la fin d'arrêt pouvant se dérouler n'importe quel jour du mois ou de l'année, il est possible de prendre en compte les périodes partiellement observées des individus en supposant que ceux-ci sont soumis à un risque homogène. Sur chaque période, l'individu est exposé à la même chance de sortie à chaque instant, c'est à dire que le taux instantané de sortie est constant : il s'agit d'un modèle exponentiel par morceau. Il est alors possible de démontrer (Rodriguez, 2010) que la vraisemblance obtenue est équivalente à celle d'une loi de Poisson. Cette approche fut déjà utilisée pour modéliser le maintien en arrêt (Fong et al., 2013) à l'aide d'un modèle linéaire généralisé.

2.3 Estimateur de Kaplan-Meier et taux bruts

2.3.1 Présentation théorique de l'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est un estimateur non paramétrique de la fonction de survie utilisable en présence de censures et troncatures. Il se caractérise graphiquement par une courbe en plateau décroissante, les points de discontinuité correspondant aux points ou un évènement se réalise, ici la sortie d'un individu. En effet, l'estimateur de Kaplan-Meier repose sur l'estimation des taux instantanés de survie, justifiant historiquement l'utilisation de table de taux pour la tarification et le provisionnement concernant l'arrêt de travail.

Nous ferons ici une présentation rapide de l'estimateur de Kaplan-Meier négligeant la troncature pour en simplifier la présentation, qui constitue une version abrégée des informations disponibles dans (Lopez, 2013). Nous rappelons comme nous l'avons fait dans la section précédente, que les observations sont composées des quantités censurées $T = \min(Y, C)$ et d'un indicateur de censure $\delta = \mathbb{1}_{Y \leq C}$.

Dans le cas d'une variable discrète, les taux instantanés de sortie et la fonction de survie peuvent s'écrire $\mu(t) = \mathcal{P}(Y = t \mid Y \geq t) = \frac{\mathcal{P}(Y=t)}{\mathcal{P}(Y \geq t)}$ et $S(t) = P(Y \geq t)$. Il vient que la fonction de survie peut se définir intégralement à partir des taux instantanés par la formule $S(t) = \prod_{t_i<t} (1 - \mu(t_i))$. Un estimateur de la fonction de survie s'obtient donc à partir d'un estimateur des taux instantanées en posant $\hat{S}(t) = \prod_{t_i<t} (1 - \hat{\mu}(t_i))$.

La prochaine étape consiste à relier les taux instantés par des quantités que nous pouvons estimer à partir des observations. Les quantités $H_0(t) = \mathcal{P}(T \leq t, \delta = 0)$, $H_1(t) = \mathcal{P}(T \leq t, \delta = 1)$ et $H(t) = \mathcal{P}(T \leq t)$ correspondent aux fonction de survie respectivement des données censurées, non censurées et des données au global. Ces trois quantités disposent d'estimateur facilement calculable qui sont $\hat{H}_0(t) = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\mathbb{1}_{T_i \leq t}$, $\hat{H}_1(t) = \frac{1}{n}\sum_{i=1}^{n}\delta_i \mathbb{1}_{T_i \leq t}$ et $\hat{H}(t) = \hat{H}_0(t) + \hat{H}_1(t)$.

Il est possible de montrer que les taux instantanés peuvent se réécrire $\mu(t) = \frac{dH_1(t)}{1-H(t^-)}$. Il en découle une estimation directe des taux instantanés et donc de la fonction de survie, l'estimateur de Kaplan-Meier s'écrivant finalement :

$$\hat{F}(t) = 1 - \prod_{i:T_i \leq t}\left(1 - \frac{\delta_i}{\sum_{j=1}^{n}\mathbb{1}_{T_i \leq T_j}}\right)$$

Sous quelques hypothèses supplémentaires, la fonction de répartition peut se réécrire sous la forme $\hat{F}(t) = \sum_{i=1}^{n} W_{i,n}\mathbb{1}_{T_i \leq t}$ avec $W_{i,n} = \frac{1}{n}\frac{\delta_i}{1-\hat{G}(T_i)}$ et $\hat{G}$ estimateur de la fonction de répartition de la censure.

On reconnaît ici la formule d'estimation usuelle d'une fonction de répartition mais pour des données pondérées par $W_{i,n}$. L'estimateur de Kaplan-Meier peut être vu comme une estimation de la fonction de survie octroyant un poids plus important aux arrêts longs puisqu'ils ont plus de chance d'être censurés. Ces poids ont tendance à exploser pour les observations non censurées les plus élevées, entraînant une certaine instabilité et une forte erreur d'approximation en fin de queue.

Les bonnes propriétés asymptotiques de l'estimateur furent démontrées par les travaux de Stute et Wang en 1993 et ceux de Gill en 1983. Ces premiers démontrèrent un résultat impliquant la consistance de l'estimateur sous l'hypothèse que $\mathcal{P}(T = C) = 0$ et que Y et C soient indépendants. Ce deuxième montra que $n^{\frac{1}{2}}\left(\hat{F}(.) - F(.)\right) \implies Z(.)$ où Z processus gaussien sous l'hypothèse que $\int \frac{dF(t)}{1-G(t^-)} < \infty$, c'est à dire qu'il n'y ait pas trop de censure en fin de queue, ce qui n'est pas nécessairement vérifié dans notre situation et peut entraîner une vitesse de convergence moindre que $n^{\frac{1}{2}}$.

Il est possible d'estimer la variance de cet estimateur, notamment avec la formule de Greenwood (Freedman, 2008). Cette possibilité ne sera pas utilisée ici, mais est un élément important de la méthode pour évaluer la marge d'erreur du modèle, qui peut apparaître par exemple dans le calcul de SCR.

Concernant l'usage de l'estimateur de Kaplan-Meier en lien avec les méthodes d'apprentissage, d'autres conditions sont nécessaires au bon fonctionnement : il convient que la censure et les variables explicatives ne soient pas corrélées. Cette condition n'est pas respectée dans notre jeu de données, la date de départ à la retraite étant directement reliée à la date de fait générateur de l'individu. Cependant, les travaux menés sur données simulées (voir II 6.3), semblent révéler que cette condition n'a finalement que peu d'impact sur les performances. De plus, il est courant, comme nous le ferons dans cette étude, de réaliser une estimation de Kaplan-Meier des durées d'arrêt pour chaque catégorie d'âge au fait générateur. Il s'agit de l'estimateur conditionnel de Kaplan-Meier, aussi appelé estimateur de Beran, qui repose sur l'hypothèse moins forte d'indépendance conditionnellement à l'âge de Y et de C, qui parait ici justifié. Le lecteur trouvera plus d'informations sur le sujet dans (Lopez, 2007).

2.3.2 Utilisation pour le calcul des taux de sortie bruts

La mise en pratique est relativement simple sous R, à l'aide du package *survival* (Terry M. Therneau and Patricia M. Grambsch, 2000) et de la fonction *survfit* permettant de réaliser directement les calculs. Nous avons alors estimé la fonction de répartition de la durée des arrêts de travail pour chacune des différentes tranches d'âge.

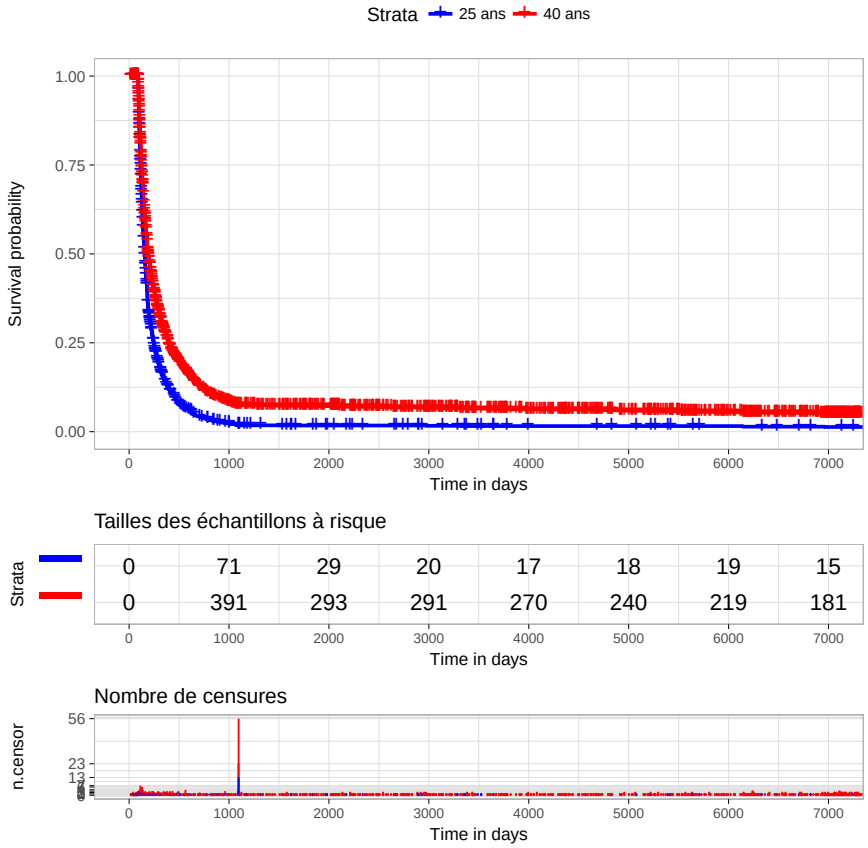


FIGURE 8: *Estimateur de Kaplan-Meier et informations annexes pour deux tranches d'âge : les individus de 25 ans (bleu) et les individus de 40 ans (rouge)*

Il est difficile de comparer les taux (la décroissance des courbes) sur la partie traitant de l'invalidité (au delà du 1096^{eme} jour). La fonction de répartition décroissant plus rapidement au départ pour les individus de 25 ans que pour les individus de 40 ans. Ces derniers doivent avoir des taux plus faibles sur cette période. L'inflexion est de même plus marquée sur la courbe bleue, révélant une décroissance plus forte des taux pour ces individus. La taille des échantillons à risque indique la troncature avant le 90^{eme} jour (aucun individu au départ), et celle-ci décroit rapidement pour ne concerner qu'une vingtaine d'individu de 25 ans au bout de 3000 jours d'arrêt, impliquant une forte volatilité des taux. Le nombre de censures est uniforme dans le temps, le nombre d'individu présent étant décroissant, cela implique que la proportion de censure augmente dans le temps, et donc une plus forte tendance à être censuré pour les arrêts longs. Le pic de censure au 1096^{eme} jour correspond aux arrêts atteignant la durée maximale légale en incapacité et n'étant pas reconduit en invalidité, que nous avons déjà évoqué.

La fonction *summary* du même package, permet d'agréger la fonction de survie pour les périodes

qui nous intéressent, en mois et année, pour enfin calculer la table des taux et la mettre sous forme de dataframe. Les valeurs manquantes de taux (fin des arrêts avant la limite possible) furent corrigées, ainsi que les taux nuls dans la base remplacée par une valeur très faible (10^{-5}) afin de ne pas géner les calculs de vraisemblance au besoin. Ces corrections ont peu d'impact sur la suite des travaux car les taux associés sont construits à partir d'un nombre réduit voire nul d'individus, alors que les méthodes de lissage utilisés sont pondérées par le nombre d'individus utilisé dans la construction des taux. Similairement aux tables du BCAC, pour tenir compte des départs à l'âge légal en retraite pour les individus et ne pas surcharger la lecture par des informations inutiles, les taux pour la partie correspondant à l'invalidité s'arrêtent à l'âge de 62 ans.

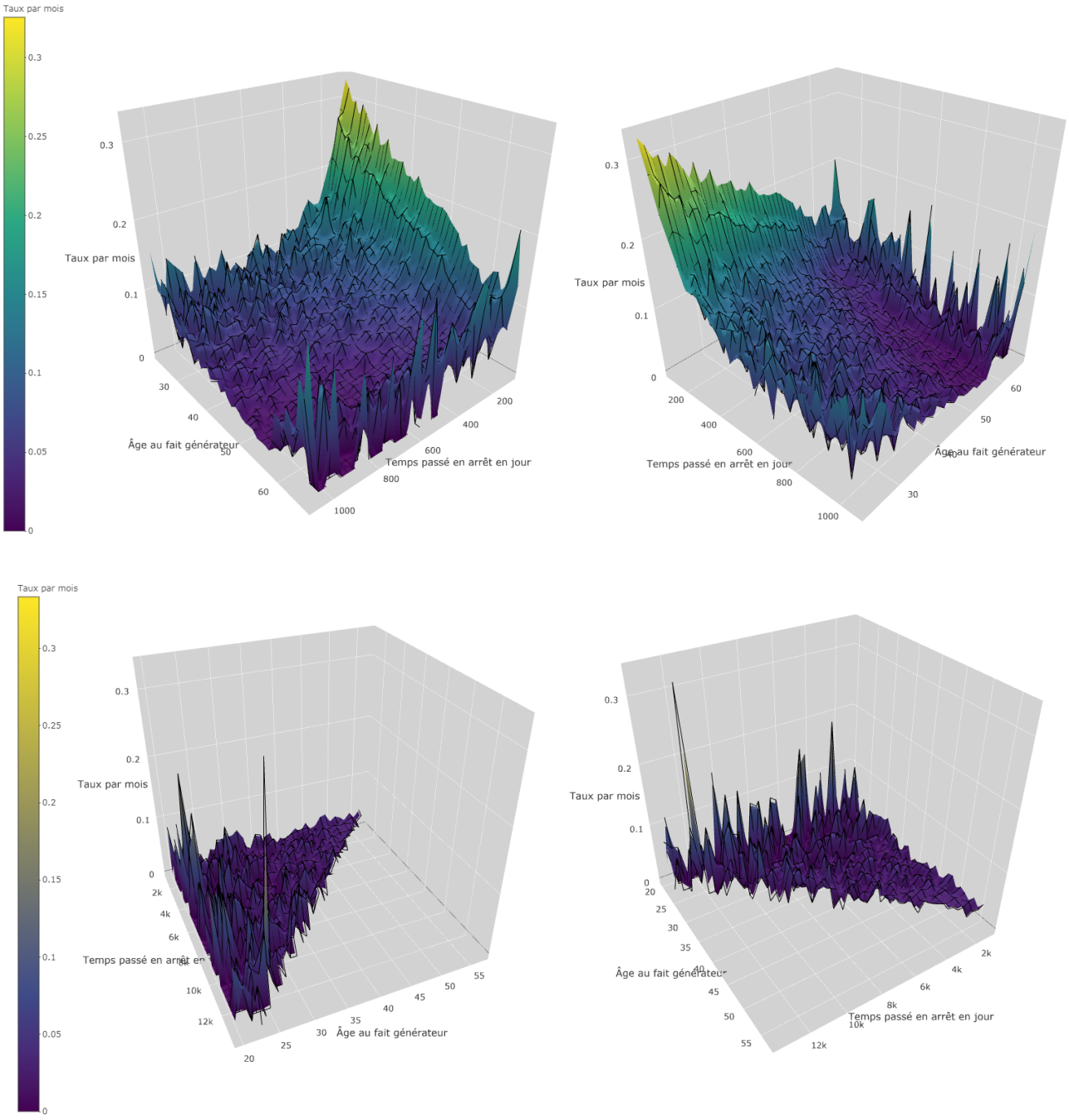


Figure 9: *Graphes des taux bruts pour la partie correspondant à l'incapacité (haut) et la partie correspondant à l'invalidité (bas) sous deux angles différents*

L'échelle des graphes étant différente, il faut rester vigilant lors de leur comparaison. Le graphe de la partie correspondant à l'incapacité est en dehors des bords (âges peu représentés) assez stable, le volume de données disponibles étant suffisant pour que l'impact de la variance de l'estimation soit négligeable vis à vis de la valeur des taux estimés. Au contraire, les taux bruts de la partie similaire à l'invalidité sont instables, et il est difficile de distinguer visuellement des tendances ou une forme de courbe, soulignant la nécessité d'un lissage sur ce segment pour obtenir une table des taux utilisable.

3 Lissage des taux

3.1 Lissage par splines : présentation

Nous présenterons dans cette partie les éléments théoriques et d'implémentation, qui permettent de replacer le lissage par splines effectué dans le cadre plus général des modèles additifs généralisés. Le lecteur souhaitant comprendre les détails théoriques de la méthode et de l'implémentation disponible sous R par le biais du package *mgcv* (Wood, 2016) trouvera toutes ces informations dans (Wood, 2006).

3.1.1 Modèle linéaire généralisé

Le modèle linéaire propose une estimation de $\mathrm{E}[Y \mid X = x]$ reposant sur l'hypothèse que le lien entre la variable à expliquer et les variables explicatives prend la forme $Y = \beta^t X + \epsilon$ où $\epsilon \sim \mathcal{N}(0, \sigma^2)$ et $\beta^t X = \sum_{i=1}^{p} \beta_i X_i$. Il repose donc sur une hypothèse de forme de la fonction minimisant l'erreur, imposant une relation linéaire entre les variables, et une hypothèse sur la forme de l'erreur, qui doit suivre une loi normale.

Notons que l'on peut réécrire le problème : $Y \mid X = x \sim \mathcal{N}(\beta^t x, \sigma^2)$ et que l'on cherche alors à déterminer le coefficient β tel que $\forall x, \mathrm{E}[Y \mid X = x] = \beta^t x$. Au lieu de chercher à déterminer cette espérance pour une loi normale, le problème peut se généraliser à toute une famille de lois. Celle-ci comprend notamment les lois dont la densité possède une structure exponentielle, c'est à dire de la forme $f(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$. θ et ϕ sont respectivement le paramètre naturel et de dispersion (en lien avec l'espérance), b et c des fonctions dépendantes de la loi, le support de cette loi restant à préciser (sous-ensemble de $\mathbb{R}$ ou $\mathbb{N}$).

Un autre ajout fut effectué afin d'englober les modèles déjà existants, notamment ceux prenant la forme $Y \mid X = x \sim \mathcal{B}\{1; u\}$ et $u = \mathrm{E}[Y \mid X = x] = F(\beta^t x)$ où F est la fonction de répartition d'une loi normale centrée réduite (modèle probit) ou d'une loi logistique (modèle logit). On utilise une fonction, nommée fonction de lien g, pour modéliser la relation entre l'espérance et les variables prédictives. Le modèle obtenu est appelé modèle linéaire généralisé, celui-ci prend la forme $Y \mid X = x \sim \mathcal{L}(\text{parametre } u)$ tel que $u = g(\mathrm{E}[Y \mid X = x]) = \beta^t x$. Un modèle linéaire généralisé peut se résumer dans le choix de la loi modélisée, la partie déterministe formée par les variables explicatives et leurs coefficients que l'on cherche à estimer, et la fonction de lien reliant la partie déterministe à l'espérance de la loi.

Notre objectif est de chercher de nouveaux taux, plus réguliers lorsque l'âge de fait générateur a et la date de début de période d varient, mais suffisamment proches des anciens. Le rapprochement des nouveaux taux $\tilde{p}_a^d$ aux anciens p_a^d est souvent mesuré à l'aide de l'erreur quadratique, ce qui s'écrit $\sum_a \sum_d \left(p_a^d - \tilde{p}_a^d\right)^2$. On reconnaît ici le programme de maximisation d'un modèle linéaire, qui pourrait être utilisé comme méthode simple de lissage des taux.

La recherche d'un effet linéaire entre l'âge et la période paraît peu adaptée : les taux lissés obtenus seraient situés sur un plan, ce qui ne semble pas correspondre à la forme du graphe obtenu concernant l'incapacité (9). Une approche possible est d'utiliser un lissage par spline, par le biais d'un modèle additif. Pour simplifier la présentation, nous omettrons les questions de pondérations, bien que le lissage des taux effectué fut pondéré par le nombre d'arrêts utilisés pour la construction des différents taux, afin d'accorder une plus grande importance aux taux construits avec plus d'arrêts donc une information plus précise.

3.1.2 Modèle additif

Un modèle additif simple adapté à notre situation est la recherche de fonctions f_1 et f_2 telles que les taux peuvent s'écrire $p_a^d = f_1(a) + f_2(d) + \epsilon_a^d$ avec les ϵ_a^d indépendant et identiquement distribués selon une loi normale centrée $\mathcal{N}(0, \sigma^2)$. Le choix de la forme des fonctions f_1 et f_2 est effectué de telle sorte à ce que le programme précédent puisse se réécrire comme un modèle linéaire, ce qui s'effectue en choisissant des fonctions prenant la forme $f(x) = \sum_{i=1}^{q} b_i(x)\beta_i$.

Dans notre cas, nous avons choisi de représenter ces fonctions sous la forme de splines cubiques : ces fonctions correspondent à des courbes, faites de sections de polynômes de degré deux joints ensembles en certains points appelés noeuds, de sorte que la fonction reste continue et que ses dérivées première et seconde le soient aussi. En utilisant q noeuds $(x_1^*, ..., x_q^*)$, la fonction se réécrit sous la forme précédente en prenant $b_1(x) = 1$, $b_2(x) = x$ et pour $i = 1, ..., q$ $b_{i+2} = R(x, x_i^*)$ avec $R(x,z) = \left[(z - \frac{1}{2})^2 - \frac{1}{12}\right]\left[(x - \frac{1}{2})^2\right]/4 - \left[|x - z| - \frac{1}{2}(|x - z| - \frac{1}{2})^2 + \frac{7}{240}\right]/24$.

En écrivant $f_1(a) = \sum_{i=1}^{q_a} b_i(a)\delta_i$ et $f_2(d) = \sum_{i=1}^{q_d} b_i(d)\gamma_i$, en notant $\beta = \left[\delta_1, ..., \delta_{q_a}, \gamma_1, ..., \gamma_{q_d}\right]^T$ et $X_i = \left[1, a_i, R(a_i, a_1^*), ..., R(a_i, a_{q_a}^*, d_i, R(d_i, d_1^*), ..., R(d_i, d_{q_d}^*)\right]$, le programme de maximisation se réécrit $\beta \in \arg\min_\beta \|y - X\beta\|^2$, ce qui correspond à la résolution d'un modèle linéaire ($\hat{\beta} = (X^T X)^{-1} X^T y$).

Plutôt que d'obtenir un impact séparé de l'âge et de la durée passée en arrêt, nous aimerions idéalement que l'effet de la durée puisse varier en fonction de l'âge. Nous utilisons pour ce faire des produits de tenseurs : l'impact de la durée est modélisé par un spline cubique, où chaque coefficient est un spline cubique dépendant de l'âge. Le problème peut ainsi s'écrire par la recherche d'une fonction f telle que $f(a,d) = \sum_{i=1}^{q_d} b_i(d)\gamma_i(a)$ avec $\gamma_i(a) = \sum_{j=1}^{q_a} b_j(a)\delta_j^i$. Il est possible de se ramener, avec un bon ordonnancement des coefficients à estimer, à un modèle similaire au précédent avec $X_i = X_{di} \otimes X_{ai}$ où $X_{di} = \left[1, d_i, R(d_i, d_1^*), ..., R(d_i, d_{q_d}^*)\right]$, $X_{ai} = \left[1, a_i, R(a_i, a_1^*), ..., R(a_i, a_{q_a}^*)\right]$ et $\otimes$ est le produit de Kronecker, d'où le terme de produit de tenseurs.

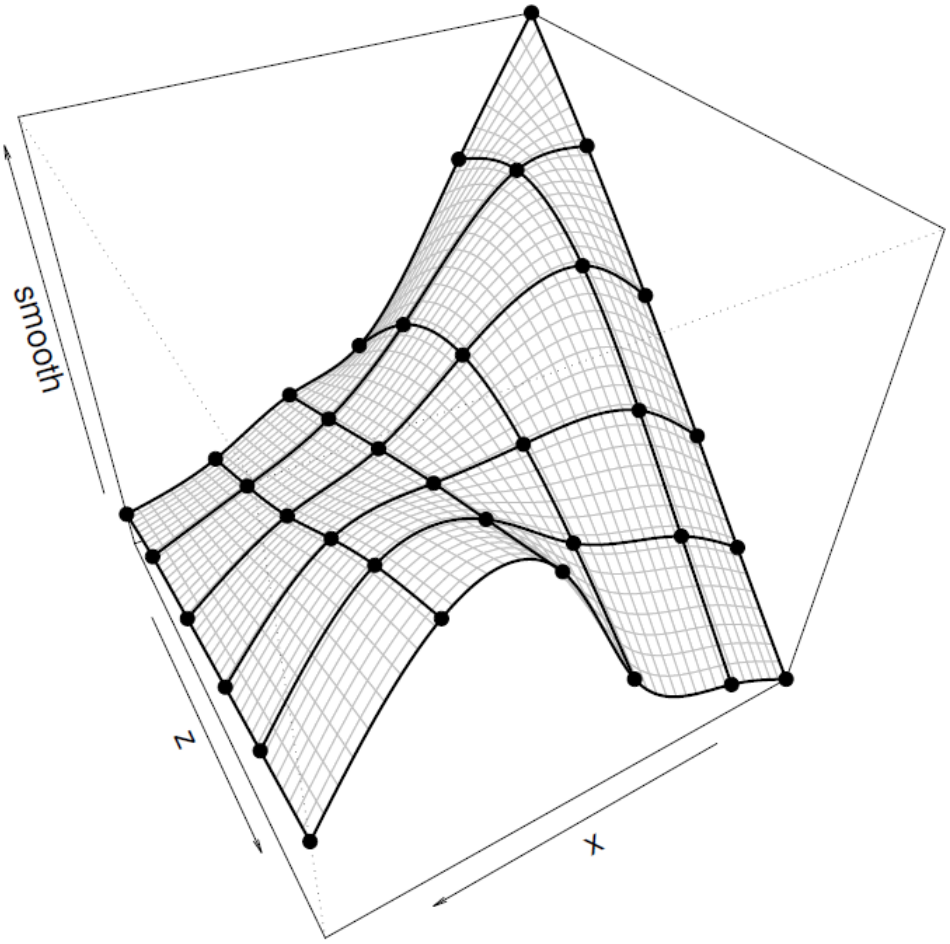


FIGURE 10: *Illustration d'un lissage par produit de tenseurs de splines cubiques, permettant que les variations vis à vis d'une coordonnée évolue lorsque l'autre coordonnée varie aussi. Source (Wood, 2006)*

Le lissage par spline cubique construit varie fortement, selon le nombre de noeuds choisi et le choix des noeuds effectué. Chercher à déterminer directement le nombre de noeuds idéal ainsi que leur position est extrêmement coûteux en temps de calcul, ainsi que difficilement interprétable. On préfère généralement choisir un nombre de noeuds suffisamment grand, espacé uniformément, pour être sûr de ne pas avoir de problème de sous-apprentissage, puis imposons un terme pénalisant l'irrégularité de la courbe afin d'éviter le sur-apprentissage. Le problème s'écrit alors comme la recherche d'une fonction f telle que $f \in \arg\min_f \mathrm{E}[(Y - f(X))^2] + \lambda \int (f''(u))^2 du$. λ est appelé coefficient de pénalisation, plus λ est élevée plus la pénalisation est importante.Pour $\lambda \to 0$ il n'y a plus de pénalisation et la fonction interpole parfaitement les données, alors que pour $\lambda \to \infty$ la fonction doit être de dérivée seconde nulle, le problème correspondant au modèle linéaire. Nous reviendrons sur le calcul de λ plus tard, et considérerons pour l'instant λ comme donné.

f étant linéaire dans les paramètres, la pénalisation peut toujours se réécrire sous une forme quadratique $\int (f''(u))^2 du = \beta^T S \beta$. Dans le cas des splines cubiques, la matrice S est telle que les deux premières lignes et colonnes sont nulles et $S_{i+2,j+2} = R(x_i^*, x_j^*)$. En notant B une racine de S (non utilisé dans les calculs, donc ne nécessite pas d'être déterminé), le problème sous forme matricielle est $\left\| \begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} X \\ \sqrt{\lambda} B \end{pmatrix} \beta \right\|^2$ et se résout encore comme un modèle linéaire. En deux dimensions, une

généralisation possible de la pénalisation qui conserve l'interprétation sur chacune des coordonnées est $\int (\frac{\partial^2 f}{\partial^2 x}(x))^2 dx + \int (\frac{\partial^2 f}{\partial^2 y}(y))^2 dy$. Cette pénalisation se réécrit encore sous forme matricielle, et ne change pas la méthode de résolution. À noter que le choix des splines cubiques prend tout son sens pour le choix d'une telle pénalisation : il est possible de montrer que les fonctions minimisant le programme précédent sont de la forme d'un spline cubique.

3.1.3 Implémentation utilisée

Afin d'obtenir le modèle additif généralisé (GAM), il suffit de transposer les outils présentés pour passer d'un modèle linéaire à un modèle additif au modèle linéaire généralisé (GLM).

Le modèle additif est suffisant pour le lissage des taux par des splines. Cependant il n'existe pas de fonction sous R pour le réaliser directement, ce qui nous aurait demandé un travail non négligeable d'implémentation des différentes fonctions. En revanche, l'ensemble des fonctions nécessaires pour réaliser facilement le modèle présenté sont disponibles pour les gams dans le package *mgcv*, et nous avons alors choisi de les utiliser pour réaliser le modèle additif correspondant comme cas particulier d'un gam. Concrètement, la différence tient dans le calcul du maximum de vraisemblance et la détermination des coefficients : le modèle additif passe par la formule présentée précédemment et le calcul de l'inversion d'une matrice, là où le gam est traité comme un glm pénalisé sur pseudo variables, la résolution du maximum de vraisemblance passant par la méthode itérative des moindres carrés repondérés itérativement pénalisés (P-IRLS, voir (Wood, 2006)), plus générale et moins précise.

Habituellement, la détermination des coefficients de pénalisations idéaux pour le lissage de tables de maintien passe par des tests d'adéquation (Planchet, 2016)). On vérifie que les taux lissés obtenus ne sont pas trop proches, ni trop éloignés, des taux d'origine. Le défaut d'une telle approche est l'incertitude qu'elle laisse sur la détermination des coefficients : plusieurs jeux de coefficients peuvent valider les critères et il n'est alors pas possible de les comparer, de définir parmi eux quel est le meilleur jeu de coefficient que l'on doit choisir.

Le package *mgcv* propose plusieurs algorithmes sophistiqués pour calculer le jeu de coefficient idéal : plusieurs mesures adaptées sont disponibles (validation croisée généralisée, maximum de vraisemblance restreint), ainsi que plusieurs méthodes plus ou moins exigeantes en temps de calcul (détermination externe au programme de maximisation des coefficients, détermination interne).

Dans le cadre de l'étude, afin que les méthodes soient facilement comparables, les coefficients de pénalisation idéaux seront traités comme les autres hyperparamètres et sélectionnés selon les performances de la table de taux obtenue sur l'échantillon de validation. Les fonctions permettant le passage du gam en table de taux, ainsi que celles pour la détermination des coefficients furent implémentées.

3.2 Lissage de Whittaker-Henderson : présentation

Le lecteur trouvera une présentation théorique un peu plus détaillée ainsi que des exemples sur le lissage de Whittaker-Henderson dans (Planchet, 2016). La présentation omettra la pondération bien qu'elle soit une fois encore utilisée.

En se limitant pour commencer à une dimension, l'objectif est la recherche de taux lissés $\tilde{p}_i$, suffisamment proche des taux bruts p_i, leur rapprochement étant mesuré par $\sum_i^n (p_i - \tilde{p}_i)^2$. Là où le modèle par spline cherchait à déterminer une forme de fonctions qui permet par la suite de calculer les taux, le lissage de Whittaker-Henderson se concentre sur la détermination des taux aux différents points. Pour que le problème ait un sens et que la solution ne soit pas $\tilde{p}_i = p_i$, nous pénalisons le programme de maximisation et imposons un critère de régularité sur les taux prenant la forme $\sum_{i=1}^{n-z}(\Delta^z p_i)^2$ où $\Delta^z p_i$ est l'opérateur de différence à l'ordre z qui s'écrit $\Delta^z p_i = \sum_{j=0}^{z} \binom{z}{j} (-1)^{z-j} p_{j+1}$.

L'importance de la pénalisation est encore prise en compte à l'aide d'un coefficient de pénalisation λ, le programme de maximisation s'écrivant $(p_i)_i \in \arg\min_{(p_i)_i} \sum_i^n (p_i - \tilde{p}_i)^2 + \lambda \sum_{i=1}^{n-z}(\Delta^z p_i)^2$.

Nous n'avons pas vu dans la littérature d'article abordant ce point de vue, mais les similitudes entre les programmes de maximisation présentés ici et pour le lissage par splines cubiques révèlent que le lissage de Whittaker-Henderson fait partie de la classe des modèles additifs. Par exemple, il est facile de montrer qu'un lissage de Whittaker-Henderson pour une régularité à l'ordre 1 et des taux uniformément espacés, correspond à un modèle additif pénalisé où les fonctions recherchées sont de la forme d'une extrapolation linéaire et le critère de pénalisation portant sur la dérivée première $f \in \arg\min_f \mathrm{E}[(Y - f(X))^2] + \lambda \int (f'(u))^2 du$.

En plus de la distinction par le choix de la base de fonction utilisée, la différence fondamentale entre le lissage de Whittaker-Henderson et un lissage par splines est la dimension de l'espace des fonctions et leur complexité : pour la construction de splines cubiques, une quinzaine de noeuds est suffisant pour couvrir la majorité des problèmes, là où le lissage de Whittaker-Henderson repose sur des formes plus simples de fonctions mais définit autant de paramètres que de modalités, ce qui correspond dans notre cas à une quarantaine de coefficients. Comme nous l'avons fait à l'ordre un, il pourrait être intéressant d'exprimer le lissage de Whittaker-Henderson sous la forme d'un modèle additif pour les autres ordres.

En écrivant l'opérateur $\Delta^z p$ sous forme matricielle $K_z p$, le lissage de Whittaker-Henderson correspond à un modèle additif où les paramètres sont reliés aux modalités et utilisant comme matrice de pénalisation $S = K_z^T K_z$. La méthode de résolution est identique, et l'on obtient : $\tilde{p} = (Id + \lambda S)^{-1} p$.

Il est possible de généraliser le lissage en deux dimensions, afin de correspondre à notre situation. La proximité des taux est alors évaluée par $\sum_a \sum_d (\tilde{p}_a^d - p_a^d)^2$ alors que, comme pour le modèle par splines, la régularité est évaluée par une régularité sur l'âge $\sum_a \sum_d (\Delta_a^z p_a^d)^2$ et une régularité sur la durée $\sum_d \sum_a (\Delta_d^z p_a^d)^2$. Le problème se réécrit comme un modèle additif pénalisé, et à condition de faire attention à la définition des matrices, la résolution est identique.

Une implémentation du lissage de Whittaker-Henderson en deux dimensions est disponible sur

le site de Frédéric Planchet. Comme pour le lissage par spline, les fonctions permettant d'effectuer la transition au format de table de taux et pour le calibrage des coefficients de pénalisation et de l'ordre z de la différence furent implémentées.

3.3 Résultats

La méthode du calcul du maximum de vraisemblance utilisée par le biais du package *mgcv*, prévue pour les modèles additifs généralisés, est inutilement trop compliquée pour la réalisation d'un lissage par splines (modèle additif). Ce désavantage est compensé par l'optimisation du package, le calibrage et l'évaluation d'un jeu de paramètres ne prenant en moyenne que 4 secondes.

Au contraire, l'implémentation manuelle (reposant sur la fonction *solve* déjà existante dans R) du lissage de Whittaker-Henderson est beaucoup plus lente, le calibrage et l'évaluation d'un jeu de paramètres prenant environ 1 minute. Cette différence s'explique par le fait que les librairies C++ de calcul matriciel de R ne sont pas optimisées. La distribution R Microsoft R Open propose de corriger ce problème. D'après nos essais sur notre ordinateur personnel, le lissage de Whittaker-Henderson est dans ces conditions 10x plus rapide, ne prenant que 6 secondes et se rapportant en ordre de grandeur au lissage par spline.

Nous n'avons malheureusement pas eu le temps nécessaire pour obtenir cette distribution, qui nécessite une installation, sur notre ordinateur de travail. Le calibrage des paramètres du lissage de Whittaker-Henderson fut donc limité, en particulier nous avons fixé l'ordre de la différence z à 2, afin d'avoir une pénalisation sur l'accélération similaire à celle employée pour la méthode des splines.

Étant donné les distinctions importantes de comportement, des jeux de coefficients distincts furent déterminés pour la partie correspondant à l'incapacité et celle correspondant à l'invalidité, le lissage de Whittaker-Henderson ne pemettant pas de toute manière de les traiter de manière jointe, les taux n'étant pas uniformément séparés (taux de mois en mois ou taux d'année en année).

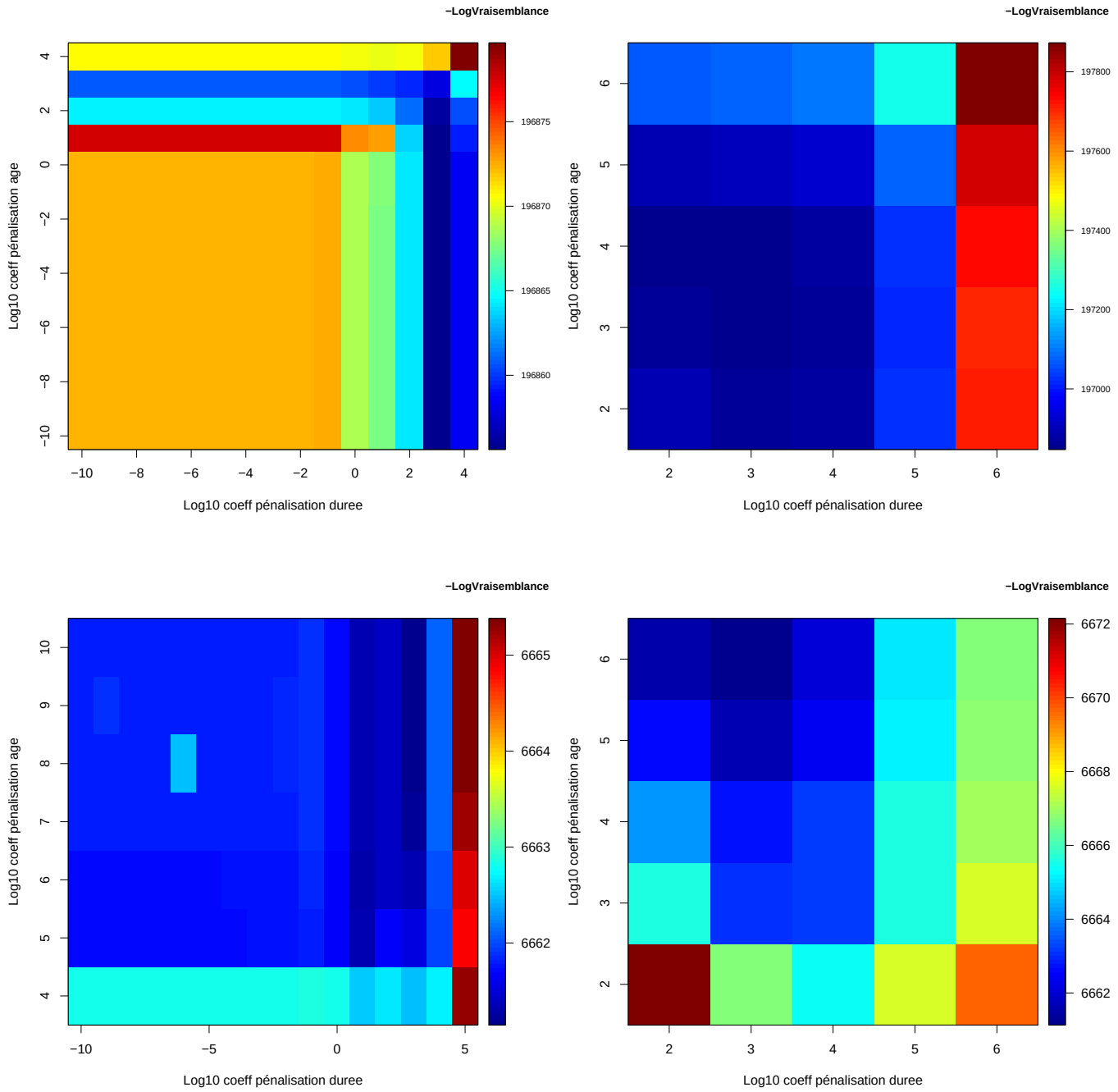


FIGURE 11: *Opposé de la logVraisemblance sur l'échantillon de validation pour le calibrage du lissage par splines (gauche) et du lissage de Whittaker-Henderson (droite), pour la partie concernant l'incapacité (haut) et l'invalidité (bas)*

Bien que les valeurs des coefficients ne soient pas directement comparables car ne reposant pas sur le même programme de maximisation, les résultats des calibrages sont similaires : les deux modèles sont optimaux pour une pénalisation moyenne sur la durée, et une pénalisation forte de l'âge concernant l'invalidité. Pour l'incapacité, le lissage de Whittaker-Henderson requiert une pénalisation moyenne de l'âge, là où la pénalisation idéale pour le lissage par spline est plus faible.

De manière générale, la vraisemblance est plus sensible aux variations du coefficient pénalisant la durée qu'à celles du coefficient pénalisant l'âge, indiquant que la durée depuis le début du sinistre contient une plus grande quantité d'information sur les taux de sortie que l'âge de fait générateur.

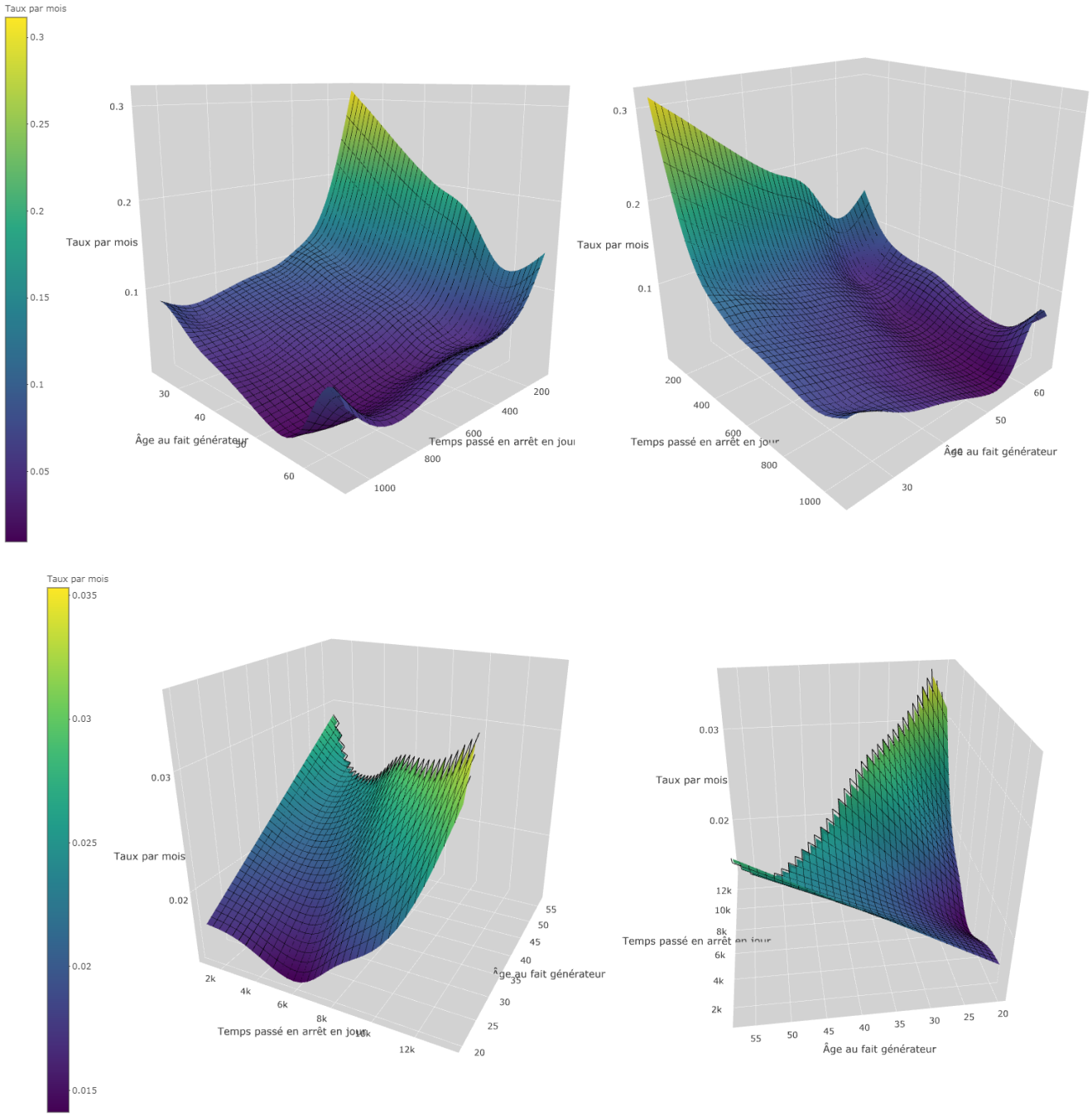


FIGURE 12: *Graphes des taux lissés par splines pour la partie correspondant à l'incapacité (haut) et la partie correspondant à l'invalidité (bas) sous deux angles différents*

La comparaison des deux graphes peut être trompeuse, puisque l'échelle n'est pas la même : pour la partie incapacité les taux ont des variations allant de 5% à 30%, les taux de la partie invalidité sont eux beaucoup plus faibles variant de 1.5% à 3.5%. On retrouve à la lecture des courbes l'impact des coefficients de pénalisation : la pénalisation est très forte sur la partie incapacité pour l'âge avec un effet quasiment linéaire, alors qu'au contraire les fluctuations des taux en fonction de l'âge sont irrégulières pour la partie incapacité.

Concernant la partie correspondant à l'incapacité, les taux de sortie sont plus élevés chez les jeunes à l'origine mais font aussi preuve d'une décroissance plus rapide, à 55 ans par exemple l'amplitude des variations des taux est beaucoup moins marquée. Cela peut s'interpréter par le fait qu'en moyenne

le temps de rétablissement augmente avec l'âge, les accidents légers des moins âgés étant regroupés en début d'incapacité, au fil du temps la proportion de sinistre grave va donc être plus forte chez les plus jeunes que chez les plus âgés expliquant la décroissance plus rapide. Une autre explication est la proportion plus importante d'accidents graves chez les jeunes (accidents de voiture), et donc un taux de décès plus important en début de courbe.

On constate une augmentation des taux pour les individus de plus de 55 ans. Les sorties pour départs à la retraite ou pour une autre cause sont difficilement distinguables sur cette tranche de population, pouvant expliquer le pic de taux en fin de table, correspondant à une qualité de données plus faible. Les taux plus élevés en milieu et début de table en revanche, sont difficilement explicables sans fortement extrapoler : on peut imaginer par exemple qu'à partir d'un certain âge les individus ont tendance à éviter les tâches les plus physiques, et donc que la nature de l'arrêt n'est pas la même (une proportion plus faible d'accident grave).

Les personnes habituées à la lecture de table d'incapacité remarqueront que, contrairement aux tables du BCAC par exemple, il n'y a pas de pic des taux de sortie en fin de table : ces pics sont dûs au passage d'incapacité en invalidité, comptabilisé comme des sorties habituellement. Dans notre situation, étant donné que nous travaillons sur la durée de l'arrêt au global, ces sorties n'apparaissent pas, rappelant que les tables de sortie d'incapacité et d'invalidité et les tables de ce mémoire ne sont pas directement comparables.

Pour la partie correspondant à l'invalidité, la lecture de fin de table révèle que les taux sont à peu près constant le long de la diagonale : en diminuant l'âge de fait générateur mais augmentant le temps passé en arrêt, les taux restent les mêmes. On en déduit qu'une des variables caractérisant la table est l'âge des individus au moment de la période concernée, les taux étant croissant en fonction de cet âge.

Le raccordement des deux courbes est aussi intéressant : on observe globalement une tendance décroissante puis croissante des taux de sortie. Cela motive nos travaux à venir qui vont tenter de modéliser directement par un mélange de lois ces deux tendances : une tendance de taux décroissante correspondant à un risque court et une tendance de taux croissante correspondant à un risque long, ce que la séparation juridique en incapacité et invalidité essaie de prendre en compte, et que nous espérons retrouver par une étude sur la durée totale des arrêts.

Le lecteur trouvera en annexe 8.C les graphes correspondant au lissage de Whittaker-Henderson. Les courbes sont similaires, les taux obtenus par le lissage de Whittaker-Henderson étant légèrement moins réguliers que ceux obtenus par la méthode des splines. À ce stade, il ne nous a pas paru pertinent de présenter une comparaison graphique des performances : les taux obtenus par spline ont une log-vraisemblance plus élevée, mais cette mesure de résultat ne dispose pas d'une interprétation concrète de cette différence, ni même sur l'importance de l'ordre de grandeur. Il n'est pas possible à partir de la comparaison des log-vraisemblances, d'affirmer que la distinction entre les deux modèles en terme de performances est importante ou non.

Néanmoins, les deux méthodes étant proches puisque construites à partir des mêmes taux bruts et les lissages étant similaires, on peut supposer que les résultats entre les deux modèles sont proches. C'est la raison pour laquelle nous avons proposé dans ce mémoire non pas une mais deux méthodes de lissage : afin de fixer un antécédent pour la comparaison des modèles à venir. On peut dire que des modèles ayant un écart en terme de performance proche de l'écart entre le lissage par spline et par Whittaker-Henderson n'a finalement que peu d'impact, et inversement.

4 Prise en compte des informations individuelles : modèle de Cox

4.1 Présentation

Les informations présentées dans ce chapitre proviennent en partie de (Saint-Pierre, 2015).

Le modèle de Cox relie les variables explicatives supplémentaires dont on cherche à modéliser l'effet $X_1, ..., X_n$ aux taux instantanés par la formule $\lambda(t, X_1, ..., X_n) = \lambda_0(t) \exp(\sum_{i=1}^{n} \beta_i X_i)$. Par exemple dans le cas de deux catégories d'individus $X_1 = 0$ et $X_1 = 1$, cela revient à considérer que les tables de taux sont proportionnelles avec comme coefficient de proportionnalité $\exp(\beta_1)$ où l'on cherche à déterminer β_1.

Il est possible de décomposer la vraisemblance totale d'un tel modèle en une partie concernant $\lambda_0(t)$ et une autre dépendante de $\exp(\sum_{i=1}^{n} \beta_i X_i)$. Dans le modèle de Cox, nous ne cherchons pas à déterminer λ_0 mais nous intéressons uniquement aux coefficients $\beta_1, ..., \beta_n$. Pour déterminer la valeur de ces coefficients, le modèle de Cox repose sur la maximisation du fragment de la vraisemblance totale concernant ces coefficients, on dit que le modèle est un modèle à vraisemblance partielle.

Pour simplifier, nous supposerons qu'à chaque instant, dans notre cas chaque jour, une seule sortie peut avoir lieu, deux arrêts ne pouvant se terminer le même jour. En réutilisant les notations introduites, T_i étant le temps de fin d'observation et δ_i l'indicateur de censure de l'individu i, la log-vraisemblance s'écrit $\sum_{i:\delta_i=1} X_i^T \beta - \log\left(\sum_{j:Y_j \geq Y_i} \exp(X_j^T \beta)\right)$. La résolution est généralement effectuée par une méthode itérative telle que la méthode de Newton-Raphson.

Nous rappelons que nous supposons que les taux dépendent de l'âge de fait générateur et s'écrivent plutôt $\lambda^a(t, X_1, ..., X_n) = \lambda_0^a(t) \exp(\sum_{i=1}^{n^a} \beta_i^a X_i)$. Plutôt que de résoudre un problème pour chaque âge, nous supposons que les coefficients β_i ne dépendent pas de l'âge. Le problème se ramène à un unique programme de maximisation identique au précédent, la distinction s'effectuant sur la somme $\sum_{j:Y_j \geq Y_i}$ qui devient $\sum_{j:Y_j \geq Y_i \wedge a_j = a_i}$.

Concernant la gestion des dates d'arrêts simultanés plusieurs méthodes existent. La plus simple est la méthode de Breslow, qui consiste à ne pas changer la vraisemblance précédente. Dans notre cas, étant donné que nous effectuons une distinction sur les différents âges de fait générateur, le nombre d'évènements simultanés est suffisamment faible pour que la méthode convienne.

L'hypothèse de proportionnalité est centrale au modèle. L'une des méthodes pour vérifier si cette hypothèse est valide pour une des variables explicative est le test des résidus de Schoenfeld. Les résidus sont calculés pour chaque date de fin d'arrêt comme la différence de la modalité prise par l'individu sorti (les variables à plusieurs modalités étant codées auparavant en variables indicatrices) et la moyenne des caractéristiques des individus à risque de sortir d'arrêt. Si l'hypothèse de proportionnalité est vérifiée, la distribution des résidus ne doit pas varier en fonction du temps. Le test des résidus consiste à effectuer un modèle linéaire sur les résidus, et à vérifier que la pente calculée ne soit pas significative. Selon la forme des variations au cours du temps, il se peut qu'un modèle linéaire ne soit pas adapté, il est donc courant en plus du test de tracer les résidus, ou s'ils sont nombreux et le graphe peu lisible, la valeur moyenne des résidus.

Habituellement, le modèle n'est effectué que si l'hypothèse de proportionnalité est valable. Dans cette étude, nous profitons de la mesure des résultats pour étudier comment se comportent les performances du modèle vis à vis de la validité de l'hypothèse de proportionnalité. À première vue, rien n'empêche qu'un modèle dont les hypothèses ne sont pas respectées d'être meilleur que les autres modèles que l'on peut construire, en particulier dans notre cas quand les modèles construits ne prennent pas en compte l'effet de toutes les variables explicatives disponibles. La modélisation relevant de l'abstraction, toute modélisation possible est inexacte, certaines étant plus fausses que d'autres. Le critère indiquant quel est le modèle le moins faux étant ici les performances sur l'échantillon de test.

Une fois les coefficients $\beta_1, ..., \beta_n$ déterminés, les taux sont calculés à l'aide de l'estimateur de Kalbfleish-Prentice. Celui-ci consiste à considérer les coefficients calculés comme fixés, et à les réinjecter dans la vraisemblance. Il est possible de montrer (Weng, 2007), que cet estimateur est équivalent à l'estimateur de Kaplan-Meier effectué sur des données pondérées, la pondération correspondant aux coefficients multiplicatifs $\exp(X^T\beta)$.

Une fois les taux bruts obtenus, un lissage est encore une fois effectué. Nous nous sommes limités à l'usage du lissage par splines cubiques présenté précédemment afin d'alléger les temps de calcul.

Il y a plusieurs possibilités quant aux choix des variables à inclure/exclure du modèle. Habituellement dans la construction de table, les variables conservées pour la réalisation du modèle de Cox sont celles passant le test de significativité. Il est commun dans la réalisation d'un glm de passer par d'autre critères, tels que les critères AIC et BIC (annexe 8.D).

La mesure de l'impact d'une variable diffère d'une méthode à l'autre, mais toutes ont le désavantage de fixer a priori une quantité d'information que doit contenir la variable : si l'information contenue dans la variable ne dépasse pas un certain seuil, celle-ci est éliminée. Rien ne garantit que le niveau d'information à retenir soit adapté au problème : un niveau trop élevé et certains effets relevant du bruit dans les données peuvent être conservés, un niveau trop bas et le modèle peut rejeter une partie de l'information pertinente disponible dans les données. Cela est équivalent à déterminer à l'avance quel est le seuil de séparation entre sous-apprentissage et sur-apprentissage. Nous préférons alors effectuer la sélection de variable en comparant les performances des différents modèles obtenus sur la base de test.

Dans notre cas, le nombre de variables, même découpées en variables indicatrices pour chacune des modalités, est restreint. Néanmoins tester l'ensemble des modèles possibles devient rapidement trop coûteux en temps de calcul, puisque $2^{(nombre\ de\ variables)}$ combinaisons différentes devraient être évaluées. Plutôt que de comparer l'ensemble des combinaisons possibles, on utilise généralement un algorithme de recherche de solution, nous utiliserons ici la méthode backward/forward (arrière/avant) (Hastie et al., 2008) qui en partant du modèle sans variable explicative l'améliore itérativement en essayant de retirer ou d'ajouter une variable.
Le pseudo-code peut s'écrire :

Fonction *sélection backward/forward*
(liste de variables prédictives) → *sélection de variables* **:**
    sélection ← toutes les variables
    modèle actuel ← modèle sur la sélection
    terminé ← Faux
    tant que *non terminé* **faire**
        terminé ← Vrai
        Phase backward :
            pour chaque *variable hors de la sélection* **faire**
                nouveau modèle de Cox sur la sélection − la variable
                calcul des taux bruts
                calibrage des coefficients de pénalisation sur la base de validation
                calcul des taux lissés
                calcul de l'erreur du modèle sur la base de validation
            fin
        end
        Phase forward :
            pour chaque *variable de la sélection* **faire**
                nouveau modèle de Cox sur la sélection + la variable
                - identique -
            fin
        end
        si *un des nouveaux modèles est meilleur* **alors**
            modèle actuel ← meilleur des modèles
            sélection ← sélection +\− variable concernée
            terminé ← Faux
    fin
    return sélection
end

Une implémentation des différentes méthodes détaillées ci-dessus sont disponibles dans le package R *survival*. Nous avons utilisé les fonctions *cox*, *survfit* et *cox.zph* permettant respectivement de réaliser le modèle de Cox, calculer les taux bruts et tester l'hypothèse de proportionnalité. Les fonctions permettant l'écriture des taux sous forme de table et le calibrage furent implémentées.

4.2 Résultats

Les variables explicatives ajoutées au modèle sont le sexe (Homme/Femme) de l'individu, sa catégorie socio-professionnelle (Cadre/Etam/Ouvrier) et la nature de la cause de l'arrêt (Privée/Professionelle). Afin de tenir compte de la différence des taux, des coefficients distincts sont déterminés pour la partie correspondant à l'incapacité et l'invalidité. Les différents tableaux regroupant l'information sur le modèle de Cox et les tests de proportionnalité sont disponibles en annexe 8.E .

La sélection de variable aboutit au même résultat qu'une sélection par significativité des coefficients pour l'incapacité, éliminant la distinction Cadre/Etam et supprimant l'information sur le sexe. Bien que la distinction Cadre/Etam ne soit pas significative pour l'invalidité, la sélection effectuée conduit à la conserver, soulignant qu'un coefficient non significatif peut contenir une information améliorant la qualité du modèle, le test utilisé étant dans ce cas trop restrictif.

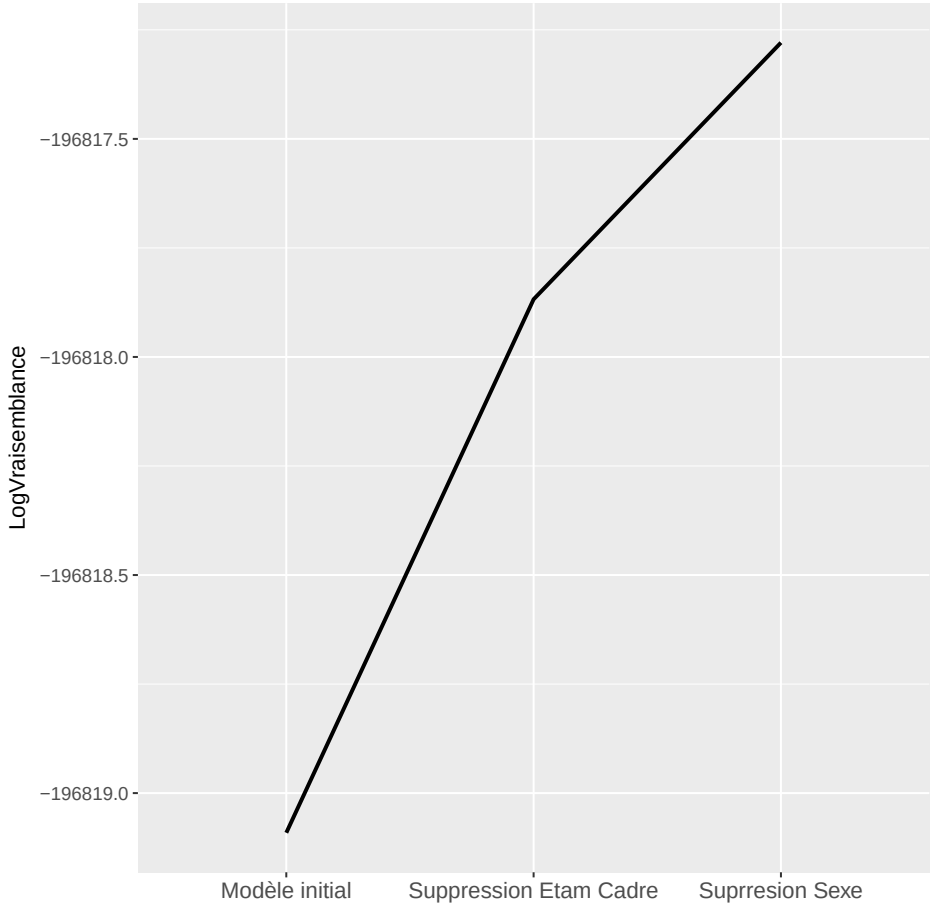


Figure 13: *Evolution de la log-vraisemblance sur la base d'apprentissage lors de la sélection de variable pour la partie incapacité*

Le calibrage des paramètres, les taux bruts et les taux lissés sont visuellement identiques à ceux obtenus par l'estimateur de Kaplan-Meier et le lissage par splines (annexe 8.F). L'unique distinction est l'ordre de grandeur des taux lissés pour le modèle de Cox sur la partie invalidité qui varient entre 0.4% et 0.9%. Cette différence s'explique par le fait que la population de référence du modèle (les cadres femmes) ne correspond pas à la population majoritaire de la base (les ouvriers hommes). Il suffit pour se ramener à un graphe d'échelle comparable d'appliquer les coefficients multiplicatifs.

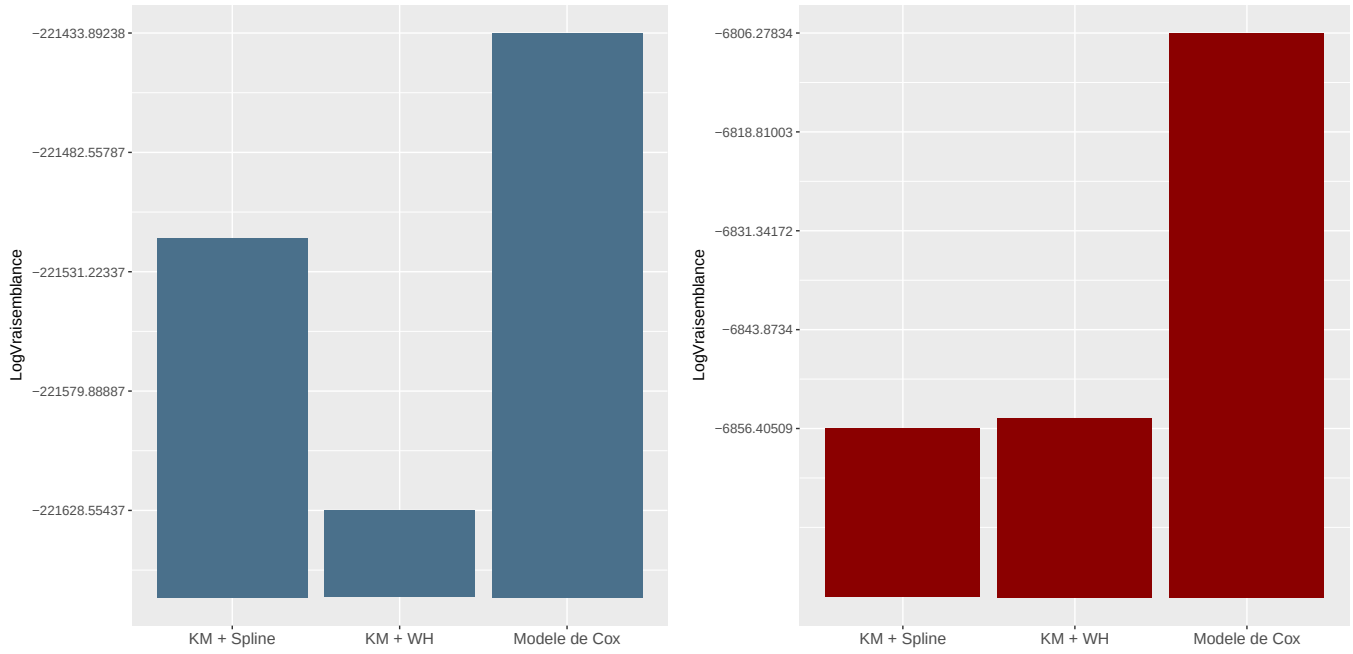


Figure 14: *LogVraisemblance des différents modèles sur la partie correspondant à l'incapacité (gauche) et l'invalidité (droite)*

Les performances du modèle de Cox sont supérieures aux deux autres modèles sur la partie incapacité et la partie invalidité. Sur la partie invalidité, l'écart de performances entre le modèle de Cox et les deux autres modèles est bien supérieur à l'écart entre les taux lissés par splines cubiques ou la méthode de Whittaker-Henderson, signe d'une amélioration significative lors de l'utilisation du modèle. À l'inverse sur la partie incapacité, les écarts de performances sont du même ordre de grandeur, révélant un faible intérêt pour l'usage du modèle de Cox sur ce segment comparé aux autres modèles.

La comparaison avec les tests de proportionnalité (annexe 8.E) confirme leur usage habituel : le gain en performance du modèle est significatif lorsque l'hypothèse de proportionnalité est respectée, et inversement. À noter que contrairement à ce que l'on aurait pu attendre, le fait que les hypothèses du modèle ne soient pas respectées n'entraîne pas de chute des performances.

Les tests de significativité reposant sur l'hypothèse de proportionnalité, il est superflu d'en discuter concernant l'incapacité. Nous avons par curiosité calibré un modèle sur l'invalidité correspondant à une sélection de variables selon leur significativité. Comparativement à l'ordre de grandeur des performances, les variations du modèle sont négligeables. Ce résultat paraît cohérent, les tests de significativité sont relativement stricts, si bien que la quantité d'information éliminée à tort est faible. Il aurait été en revanche intéressant de détecter le phénomène inverse où la significativité des coefficients aboutit à la sélection à tort d'une variable superflue, pour savoir si dans ces situations le gain de performance est significatif ou non.

Deuxième partie

Modèles de durée et arrêt de travail

5 Présentation des réseaux de neurones

5.1 Présentation théorique

Les réseaux de neurones s'inspirent du fonctionnement des neurones biologiques, un neurone transmettant une information de sortie suite à l'information reçue, information modulée par le biais de synapses. On supposera que la modulation s'effectue par l'attribution d'un poids à l'information, et que l'information de sortie (output) est une fonction non-linéaire de l'information d'entrée (input). Un neurone peut donc être vu comme une fonction $f : \mathbb{R}^{+1} \times \mathbb{R}^p \to \mathbb{R}, f(W, x) = g(\sum_{i=i}^{p} w_i x_i + w_{p+1})$ où x est le vecteur d'inputs, W les poids associés et g la fonction non linéaire utilisée appelée fonction de transfert.

Les neurones sont ensuite organisés en un réseau cohérent, de telle sorte que les outputs de certains sont les inputs d'autres. Nous nous limiterons dans notre étude à la construction de perceptron à une ou plusieurs couches cachées : la première couche correspond aux inputs du modèle et la dernière à ses outputs, les connexions entre les neurones étant telles que tous les neurones d'une couche à l'autre sont connectés

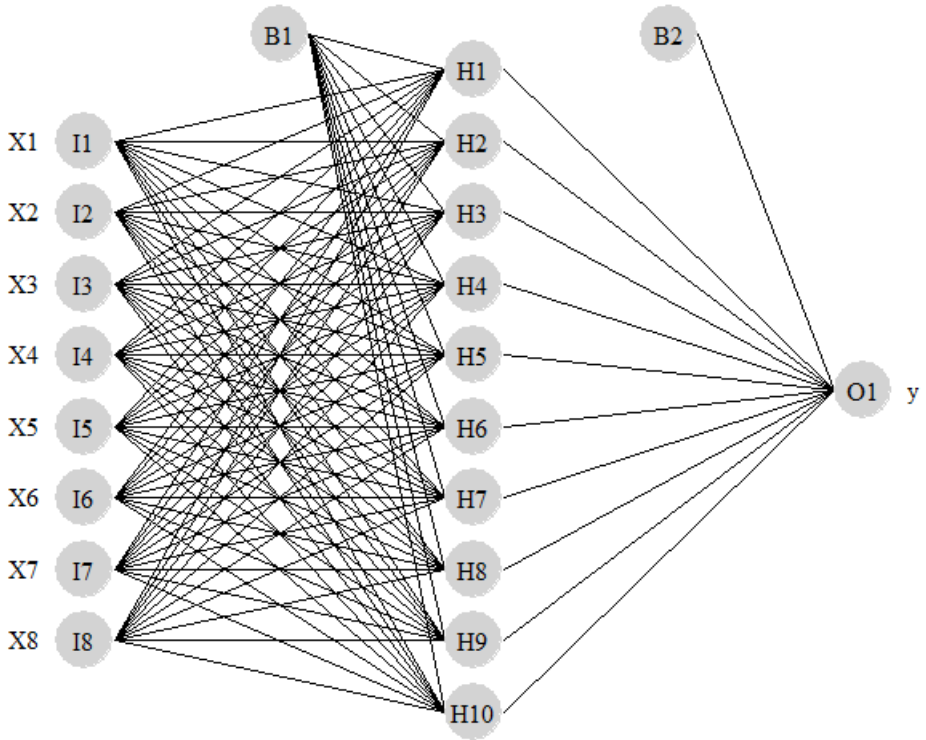


Figure 15: *Exemple de représentation d'un réseau de neurones à une couche cachée de 10 neurones et 8 variables prédictives pour la prédiction d'un unique output*

L'ensemble des informations présentées et utilisées pour ce mémoire concernant les réseaux de neurones, et beaucoup d'autres encore, sont disponibles dans le livre (Goodfellow et al., 2016).

5.1.1 Graphe orienté et descente de gradient

Il est commun de représenter les réseaux de neurones sous la forme de graphe orienté, par lisibilité et pour la compréhension de l'algorithme d'optimisation employé. On peut définir un graphe comme composé de noeuds représentant une variable, pouvant prendre de nombreuses formes (scalaire, vecteur, matrice, tenseur) et d'opérations reliant les noeuds les uns aux autres, les noeuds provenant soit du résultat d'une opération ou de données et paramètres du modèle. Sans chercher à faire de définition rigoureuse, on parle de graphes orientés lorsqu'il n'est pas possible de retourner à un noeud en parcourant le graphe. L'apprentissage du modèle consiste en la détermination de certains noeuds de paramètres, afin de minimiser la valeur en sortie.

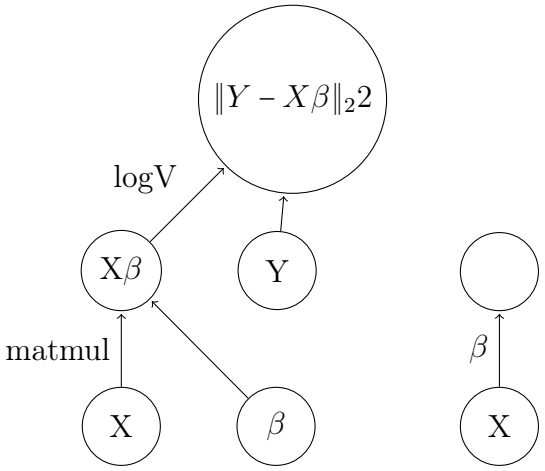


Figure 16: *Exemple d'un réseau de neurones (gauche) sous la forme d'un graphe orienté pour la résolution d'un modèle linéaire par la méthode des moindres carrés. X est la matrice des inputs à n lignes (individus) et k colonnes (variables explicatives), β la matrice k,1 des coefficients à estimer et Y la matrice n,1 de la variable expliquée. Le programme de minimisation des moindres carrés s'écrit ici comme la recherche de la valeur d'un noeud (β) minimisant la valeur en sortie du graphe. Dans la suite de mémoire, nous préférerons une écriture condensée (droite) pour plus de lisibilité*

Une telle écriture facilite grandement l'usage d'algorithme de direction de descente pour le programme d'optimisation. Parmi les méthodes les plus simples de cette classe, nous trouvons la descente de gradient à pas constant : celle-ci consiste à rechercher le minimum d'une fonction en effectuant des pas de taille fixes dans la direction de plus grande descente qu'est le gradient.

Descente de gradient à pas constant ϵ pour la recherche de $x \in \arg\min_x f(x)$
Initialisation de x
tant que *critère d'arrêt non atteint* **faire**
  Calcul du gradient $\Delta_x f(x)$
  Pas dans la direction : $x = x + \epsilon \Delta_x f(x)$
fin
Retourne x

Après chaque pas effectué pour les paramètres du modèle que l'on cherche à optimiser, il est facile de mettre à jour les autres noeuds en parcourant le graphe orienté, cette phase de mise à jour du modèle est appelée phase forward.

Alors que le gradient pour les derniers noeuds du graphe se calcule rapidement, il est rare de disposer d'une formule directe pour les noeuds plus en amont. La règle de chaîne de calcul, pouvant s'écrire $\Delta_x z = \left(\frac{\partial y}{\partial x}\right)^T \Delta_y z$ en notant $y = g(x)$, $z = f(y) = f(g(x))$ et $\frac{\partial y}{\partial x}$ la matrice jacobienne de g, nous permet néanmoins de calculer la valeur du gradient en un noeud à partir de la valeur du gradient des noeuds fils, des opérations utilisées pour l'obtention des noeuds fils et de la valeur des noeuds. La valeur du gradient en tout noeud du graphe s'obtient ainsi récursivement, en remontant au fur et à mesure le graphe, on parle alors de technique de rétropropagation. La phase de réalisation de ces calculs dans l'algorithme de descente porte le nom de phase backward.

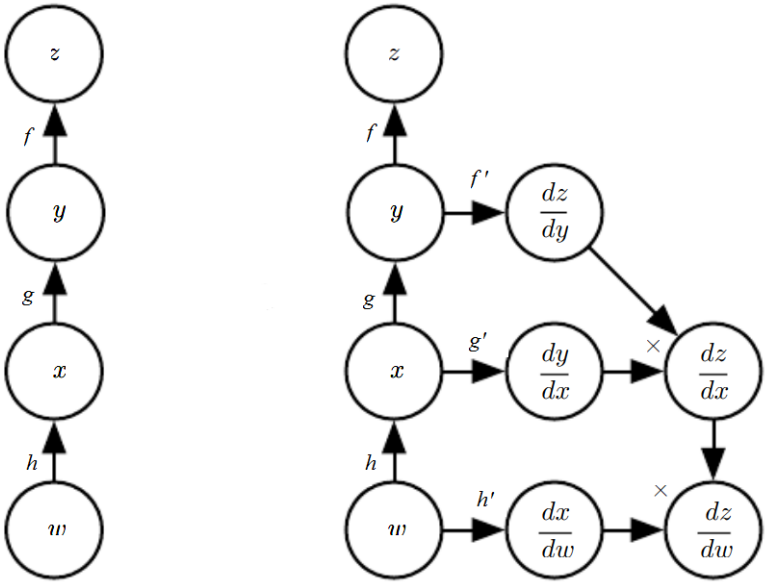


FIGURE 17: *Phase forward (gauche) et phase backward (droite) de la descente de gradient à l'aide de la technique de rétropropagation. Notons que le parcours dans le sens opposé du graphe orienté s'écrit lui aussi sous la forme d'un graphe orienté. Source (Goodfellow et al., 2016)*

Un des intérêts des réseaux de neurones réside dans la facilité d'implémentation des calculs grâce à la technique de rétropropagation. Dans notre cas les calculs de la phase de rétropropagation sont écrits manuellement, mais des architectures de Deep Learning (comme Tensorflow) sont conçues pour tirer parti de la représentation sous forme de graphe, et construisent automatiquement le graphe utilisé dans la phase backward à partir du graphe de départ.

5.1.2 Spécificité des modèles de réseaux de neurones

Il est possible de réaliser par le biais de réseaux de neurones des modèles additifs généralisés, de manière assez similaire à ce que nous avons présenté dans le cadre du modèle linéaire (voir graphe 16). Nous rappelons qu'un modèle additif généralisé n'est rien d'autre qu'un modèle linéaire généralisé sur une transformée de variables, le modèle linéaire généralisé n'étant lui même qu'un cas de modèle linéaire pour lequel la fonction à maximiser n'est plus celle des moindres carrés. Les réseaux de neurones permettent cependant l'utilisation de modèle plus complexe.

Là où la maximisation des modèles additifs généralisés ne concerne qu'un unique paramètre de la loi (l'espérance) relié aux inputs, il est possible avec les réseaux de neurones de maximiser la vraisemblance pour plusieurs paramètres simultanément.

Pour cerner le fonctionnement classique d'un perceptron multi-couches, il est commun de le séparer en deux parties : la liaison entre la dernière couche cachée et les outputs relie les paramètres de la loi à une transformée non linéaire d'une combinaison linéaire des outputs de la couche, alors que les opérations entre les inputs et la dernière couche cachée de neurones correspond à une transformation des variables. Un perceptron multi-couches revient à effectuer une transformée des variables, puis à relier les paramètres de la loi à cette transformée par le biais d'une transformation non-linéaire d'une combinaison linéaire, comme dans un modèle additif généralisé. Pour un gam la transformée de variables est considérée fixe et l'apprentissage a lieu uniquement sur le modèle linéaire, mais pour un perceptron il est possible d'optimiser simultanément les poids du modèle linéaire et la transformation des variables explicatives.

Alors que pour un gam l'utilisateur se doit de rechercher une base de fonctions adaptée en amont pour se rapprocher du véritable impact des variables explicatives (quelle fonction de lien, quelle forme pour quelle variable, quelles intéractions modélisées), le modèle le détermine de lui-même pour un perceptron. Il a en effet été prouvé (voir annexe (Dupré, 2004)) qu'un perceptron à une couche cachée est capable de modéliser n'importe quelle intéraction lorsque le nombre de neurones tend vers l'infini.

L'absence de forme prédéfinie peut améliorer le modèle en se rapprochant de la véritable relation entre les variables explicatives et les paramètres de la loi mais présente des inconvénients. Comparé à un modèle où la transformée de variable est définie en amont et se rapproche de la véritable nature des intéractions, rien ne garantit que l'apprentissage du modèle sans forme prédéfinie aboutira à des performances supérieures. De plus, la définition précise de la transformée de variables choisie par le modèle est souvent hors de portée, l'utilisateur a donc toujours à effectuer un arbitrage entre la perte d'interprétabilité du modèle et le gain potentiel sur les performances.

Pour autant, sans rentrer dans les détails, plusieurs outils d'interprétation similaires à ceux utilisables pour les forêts d'arbres d'apprentissage (Random Forest, Gradient Boosting Machine, Extreme Gradient Boosting) sont disponibles :

1. Pour un modèle suffisamment petit l'étude des poids des modèles est faisable, pour les modèles plus larges la distribution des poids pour chacune des couches cachées restant disponible
2. Comme pour tout modèle, les effets marginaux moyens d'une ou plusieurs variables sont calculables
3. Il est possible de mesurer l'importance relative de chaque variable explicative vis à vis des différents paramètres du modèle (voir (Garson, 1991))

5.2 Implémentation utilisée

Nous avons choisi pour ce mémoire de réaliser un package de réseaux de neurones sous R bénéficiant d'une infrastructure en C++ afin d'optimiser les temps de calculs. La réalisation d'un tel package nous assure un contrôle sur les calculs réalisés et la possibilité de personnaliser le programme d'optimisation. En particulier, il est nécessaire de sélectionner la vraisemblance à maximiser et les paramètres du modèle pour réaliser les modèles utilisés par la suite.

À postériori, nous recommandons l'usage d'une architecture existante telle que Tensorflow pour la réalisation de réseaux de neurones : bien que la prise en main soit plus complexe et le contrôle de l'utilisateur plus faible qu'une implémentation artisanale, de nombreuses fonctionnalités importantes et coûteuses à implémenter manuellement y sont déjà disponibles.

La construction du package fut effectué sous Rstudio qui permet par la gestion de projet de travailler sur les différents fichiers constituant le package et d'écrire dans une même interface le code R, le code C++ et les fichiers de présentation.

Il dispose de plusieurs fonctionnalités adaptées à l'implémentation de package avec la possibilité de détacher, installer et charger rapidement le package en construction pour le tester. Il possède aussi une interface de jonction avec git (logiciel utile à la gestion des versions d'un projet), et plusieurs éléments pour générer automatiquement des documents nécessaires au package comme la liste des fonctions à mettre à disposition des utilisateurs.

La jonction entre R et C++ est de base difficile, l'utilisateur devant gérer le transfert des informations entre les deux interfaces et faire attention à ce que les objets R sous C++ (type SEXP) soient correctement protégés pour éviter d'être considérés comme une information superflue en mémoire et automatiquement supprimés par R.

Le package *Rcpp* (Eddelbuettel, 2013) propose de résoudre ces problèmes à la place du programmeur : il suffit d'indiquer les fonctions tournant en C++ à exporter sous R pour que le package réalise les conversions et les protections nécessaires. Par le biais de Rcpp sugar, celui-ci met aussi à disposition de nouvelles classes et fonctions facilitant grandement l'écriture en C++. On en retient notamment les types $Rcpp :: vector$ et $Rcpp :: list$, équivalent en terme d'utilisation des types vecteur et liste disponibles sous R. *RcppArmadillo* (Eddelbuettel, 2013) est une extension du package Rcpp, automatisant les conversions des classes définies dans le package Armadillo, donnant accès à un grand nombre d'opérations et fonctions matricielles.

Nous avons implémenté le type "réseau de neurones" sous R comme des objets S3, ce qui est la manière la plus simple et la plus commune de définir une nouvelle classe d'objet sous R. Par exemple, la fonction *predict* vérifie le type de l'objet défini lors de sa construction et applique la méthode S3 conçue pour ce type d'objet. Il existe ainsi une méthode *predict* pour les objets de type lm (modèle linéaire) et pour les objets de type gbm (gradient boosting machine). Une autre manière de définir de nouveaux types sous R est par le biais d'objets S4. Cette méthode est plus restrictive, mais l'implémentation est plus rigoureuse et plus proche d'autres langages de programmation comme le C++.

Le package consiste en un perceptron à une ou plusieurs couches cachées dont le programme d'optimisation repose sur une descente de gradient à pas constant et sans momentum. Aucune règle d'arrêt n'est définie en interne mais une fonction fut implémentée afin de rajouter des itérations à un réseau de neurones existant. Cette fonction est utilisée pour regarder régulièrement les performances du modèle sur un échantillon de validation, afin d'arrêter l'apprentissage lorsque les performances cessent de s'améliorer. On parle alors d'early stopping, la méthode permettant de s'assurer que le modèle apprenne l'information pertinente des données mais ne recueille pas le bruit dans les données, évitant le sous-apprentissage et le sur-apprentissage.

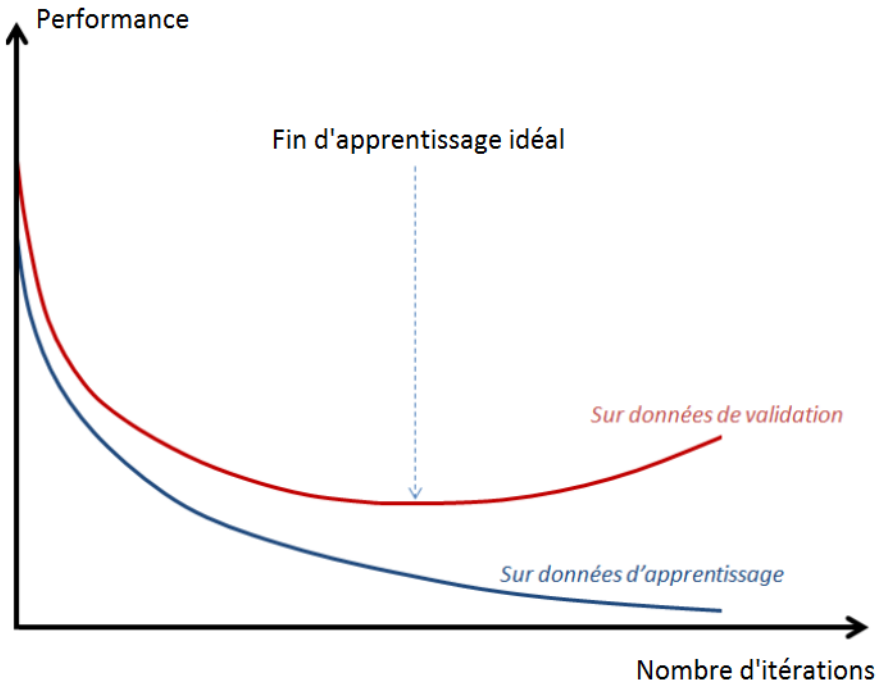


FIGURE 18: *Illustration de la méthode d'early stopping*

Le package permet d'utiliser une pondération pour les observations et propose plusieurs programmes de maximisation (différentes lois et donc paramètres de lois).

L'initialisation des poids fut effectuée comme conseillé dans (Goodfellow et al., 2016). Pour les biais, le biais de sortie est initialisé à la valeur moyenne de sortie, celui des autres couches est nul. La matrice de poids allant de m inputs à n outputs est initialisée par un tirage uniforme entre $-\frac{1}{\sqrt{m+n}}$ et $\frac{1}{\sqrt{m+n}}$. On estime cette initialisation comme un bon arbitrage entre le risque d'éloignement à la solution idéale et la distinction de l'importance des différents neurones de la couche (afin d'éviter que la descente soit identique pour tous les neurones d'une couche).

La fonction d'activation du neurone de sortie est l'identité (classique en régression) et la fonction d'activation des neurones des couches cachées est la fonction ReLU ($f(x) = 0$ si $x < 0$ et $f(x) = x$ sinon) pour un modèle à plusieurs couches et la fonction sigmoïde ($f(x) = \frac{1}{1+e^{-x}}$) pour une unique couche cachée. Les travaux récents montrent l'avantage de l'utilisation d'une fonction ReLU (notamment pour le gain sur le calcul du gradient) pour des réseaux profonds, mais nous pensons que la fonction sigmoïde permet une transformation plus forte dans le cadre d'un perceptron à une couche et est donc à favoriser.

Afin d'accélérer les temps de calcul, la méthode stochastic batch (tirage d'un sous-échantillon

aléatoire à chaque itération sur lequel est effectué la descente de gradient) est implémentée. De plus, une telle méthode permet d'ajouter une perturbation aléatoire au gradient, ce qui peut avoir des bénéfices pour éviter la convergence vers un minimum uniquement local.

Fonction *Apprentissage du modèle*
(données, taille de batch b, nombre d'itérations n, modèle) → *modèle* **:**
  pour $i = 0$ *à* $n - 1$ **faire**
    sélection aléatoire d'un échantillon de taille b parmi les données
    phase forward pour le calcul des noeuds du graphe sur l'échantillon
    phase backward pour le calcul du gradient sur l'échantillon
    mise à jour des poids du modèle
  fin
  renvoi modèle
end

Il aurait été intéressant de connaître le gain en vitesse de calcul de l'utilisation de C++ vis à vis d'une implémentation intégralement sous R. Malheureusement, à notre connaissance aucun package R ne possède le même programme d'optimisation (descente de gradient simple à pas constant et arrêt par early stopping) et nous n'avons pas pu faire de comparaisons. Selon les benchmarks sur d'autres projets du package RcppArmadillo, on s'attend néanmoins à ce que cette implémentation soit environ 5 à 6 fois plus rapide qu'une implémentation intégralement sous R.

6 Comparaison de méthodes d'adaptation à la censure et la troncature

Afin d'utiliser les modèles de durée pour des problèmes où la censure et la troncature sont des enjeux importants, il est nécessaire de les adapter. Dans cette partie, nous comparons deux de ces techniques d'adaptation sur des données simulées et des données réelles : une approche classique s'inspirant de l'estimateur de Kaplan-Meier et une approche plus moderne avec une modification de la fonction à maximiser. Cette comparaison est aussi l'occasion de voir une utilisation de l'implémentation des réseaux de neurones effectuée.

Nous rappelons que les résultats de cette comparaison dépend de l'implémentation effectuée, du programme d'optimisation choisi et des données simulées et réelles utilisées.

6.1 Présentation des méthodes et résultats attendus

6.1.1 Pondération des données non censurées

Comme nous l'avons vu en partie I 2.3, la fonction de répartition obtenue par l'estimateur de Kaplan-Meier est construite à partir des données non-censurées que l'on pondère. Cette pondération peut être perçue comme une volonté de reconstruire les données pour simuler l'absence de censure en leur attribuant un poids correspondant à l'inverse de leur probabilité d'être censurée (Inverse

Probability Censoring Weights). Par exemple, une observation ayant 50% de chance d'être censurée doit être comptée deux fois pour simuler une absence de censure.

Une première technique d'adaptation est finalement d'utiliser les données non-censurées pondérées par l'inverse de leur probabilité d'être censurée, obtenu par l'estimateur de Kaplan-Meier, pour l'apprentissage des modèles. Le lecteur trouvera plus de détails et la preuve des bonnes propriétés asymptotiques de la technique dans le cadre des d'arbres d'apprentissage dans (Lopez et al., 2015).

6.1.2 Vraisemblance totale

Les méthodes d'apprentissage supervisées reposent sur la minimisation d'une fonction de perte sur une base d'apprentissage, mesurant l'écart entre la prédiction et la valeur cible. Les améliorations de ces dernières années dans le domaine reposent entre autre sur l'utilisation de fonctions de perte en lien avec la maximisation de la vraisemblance pour une loi choisie.

Une idée naturelle est alors de chercher à maximiser la vraisemblance sur l'ensemble des données censurées, en notant δ l'indicateur d'absence de censures et τ la troncature, celle-ci peut s'écrire dans le cas de données indépendantes identiquement distribuées censurées à droite et tronquées à gauche :

$$V_\theta(Y_1, ..., Y_n) = \prod_1^n \mathcal{P}_\theta(Y_i)^{\delta_i} \mathcal{S}_\theta(Y_i)^{1-\delta_i} \mathcal{S}_\theta(\tau_i)^{-1}$$

Par le passé, l'absence de résolution analytique du programme de maximisation a freiné l'utilisation de telle méthode. Désormais, avec la puissance de calcul et les outils disponibles, une résolution par descente de gradient facilement implémentable est possible.

6.1.3 Conjectures entre les deux méthodes

On s'attend à ce que les méthodes aient les avantages/inconvénients suivants :

1. La méthode par pondération est extrêmement facile d'utilisation, l'ensemble des méthodes d'apprentissage sous R par exemple propose d'indiquer en argument la pondération à utiliser. À l'inverse, la méthode par vraisemblance nécessite de pouvoir modifier le programme de maximisation ainsi que les paramètres des modèles.
2. De plus, l'absence de résolution analytique pour la méthode par vraisemblance restreint son utilisation à certaines méthodes d'apprentissage. Les méthodes d'arbres simple notamment, nécessiteraient un temps de calcul incroyable afin de calculer la valeur du noeud à chaque coupure envisagée par une méthode de type Newton-Raphson. Une personne souhaitant utiliser cette correction pour des méthodes d'arbres doit alors se restreindre à certains algorithmes adaptés (des méthodes de boosting pour les arbres par exemple).
3. La méthode par pondération peut être vue en deux étapes, une étape pour obtenir la pondération et une étape de maximisation là où la méthode par vraisemblance générale n'en nécessite qu'une seule. En particulier, les bonnes propriétés de la pondération et de la maximisation sont uniquement asymptotiques. On s'attend alors à ce que la méthode par pondération se dégrade

plus rapidement lorsque le nombre de données diminue ou le pourcentage de données censurées augmente (nous ne testerons que la seconde hypothèse).

4. La pondération étant sensible aux valeurs extrêmes, on s'attend à ce que la méthode par pondération se dégrade plus rapidement lorsque le pourcentage de valeurs extrêmes augmente.
5. La méthode des poids permet de corriger la censure, mais est biaisée en présence de troncatures constantes. Par exemple, supposons que les données proviennent d'une loi normale de moyenne 0.5 et écart-type 1 tronquée en 0. Déterminer une loi normale sur les données obtenues n'a alors rien à voir avec la recherche d'une loi normale tronquée en 0. En présence d'un seuil de troncature, la méthode par pondération doit alors être remplacée par une méthode mixte de pondération avec une fonction de perte tenant compte de ce seuil.
6. Les bonnes propriétés asymptotiques de la méthode de pondération reposent sur les mêmes hypothèses que Kaplan-Meier, entre autre l'indépendance entre la censure et les variables explicatives (covariables). On s'attend alors à une dégradation des performances de la méthode par pondération en cas de corrélation des covariables et de la censure, là où la méthode par vraisemblance générale n'est pas affectée. Une "solution" envisageable au problème est d'utiliser un Kaplan-Meier conditionnel (effectuer une pondération séparée pour chaque valeur possible de la variable supposée être corrélée à la censure). Cela nécessite cependant une variable qualitative, et implique un éclatement des données disponibles pour la pondération, dégradant les performances comme envisagé dans les points 4 et 5.

6.2 Implémentation des méthodes

Afin de perturber au minimum la comparaison, celle-ci doit être effectuée pour une méthode d'apprentissage et une implémentation communes : une fois avec une maximisation de la vraisemblance sur les données non censurées et pondérées, une fois avec maximisation de la vraisemblance sur l'ensemble des données censure comprise.

Nous nous intéresserons par la suite à la maximisation de la vraisemblance d'une loi normale. Un tel modèle permet de souligner que même en l'absence de résolution analytique (pas de formule pour la fonction de répartition d'une loi normale), l'apprentissage est possible. La maximisation de la vraisemblance d'une loi normale peut être vue de plus comme une extension de la méthode des moindres carrés, méthode qui fut utilisée dans des travaux antérieurs (Lopez et al., 2015).

Dans la suite de cette partie, nous nous contenterons d'étudier l'évolution de l'erreur des méthodes au fur et à mesure des itérations sur la base d'apprentissage et une base de test, ce qui est suffisant pour en effectuer la comparaison.

Notre package comporte entre autres la minimisation des moindres carrés et de la vraisemblance d'une loi normale. Nous rappelons qu'avec la troncature, la maximisation de l'espérance et de la variance ne sont plus séparables, le réseau de neurones devant alors renvoyer les deux paramètres, moyenne et variance.

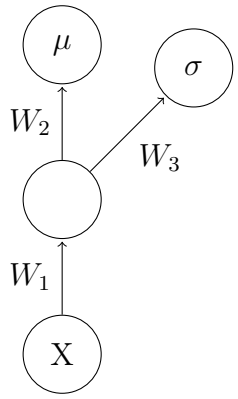


FIGURE 19: *Forme du réseau utilisé pour la loi normale*

Le calcul des poids pour la méthode par pondération n'est pas disponible pour l'ensemble des cas de figures. Le package R pec (Mogensen et al., 2012) que nous avons utilisé permet de calculer les poids dans le cadre d'une censure à droite, mais ne gère pas une troncature à gauche. Cela ne pose pas de problèmes pour l'étude sur données simulées, mais malgré une correction de la vraisemblance pour une troncature constante à gauche pour chaque individu (minimum de 90 jours de carence), les poids construits sont biaisés sur données réelles, défavorisant la méthode lors de cette comparaison.

Dans le cas de données censurées à droite avec troncature à gauche τ, la fonction à minimiser (opposé de la logvraisemblance) s'écrit, en notant Φ et ϕ les fonctions de répartition et de densité respectivement d'une loi normale centrée réduite :

$$F = -LogV = \sum_{i=1}^{n} \delta_i \left[\frac{(Y_i - \mu)^2}{2\sigma^2} + \ln(\sigma) \right] - (1 - \delta_i) \ln\left(1 - \Phi\left(\frac{Y_i - \mu}{\sigma}\right)\right) + \ln\left(1 - \Phi\left(\frac{\tau_i - \mu}{\sigma}\right)\right)$$

Les dérivées par rapport à μ et σ de la fonction de perte sont alors :

$$\frac{\partial F}{\partial \mu} = \sum_{i=1}^{n} \delta_i \frac{\mu - Y_i}{\sigma} - (1 - \delta_i) \frac{\frac{1}{\sigma}\phi\left(\frac{Y_i - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{Y_i - \mu}{\sigma}\right)} + \frac{\frac{1}{\sigma}\phi\left(\frac{\tau - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\tau - \mu}{\sigma}\right)}$$

$$\frac{\partial F}{\partial \sigma} = \sum_{i=1}^{n} \delta_i \left[\frac{-(Y_i - \mu)^2}{\sigma^3} + \frac{1}{\sigma} \right] - (1 - \delta_i) \left[\frac{Y_i - \mu}{\sigma} \frac{\frac{1}{\sigma}\phi\left(\frac{Y_i - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{Y_i - \mu}{\sigma}\right)} \right] + \frac{\tau_i - \mu}{\sigma} \frac{\frac{1}{\sigma}\phi\left(\frac{\tau_i - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\tau_i - \mu}{\sigma}\right)}$$

6.3 Test des conjectures sur données simulées

Les données simulées consistent en une variable cible et trois variables explicatives. Des effets linéaires et non linéaires (carré,cosinus) sont simulés reliant les paramètres du modèle aux covariables. Le nombre d'observations fut de l'ordre du millier, et les paramètres des réseaux (taille des batchs, pas de l'apprentissage) furent identiques pour les modèles comparés et sélectionnés afin de bénéficier de temps de calcul faible tout en s'assurant de la convergence des algorithmes.

Dans la première partie, plusieurs types de lois (normale variance homogène, normale variance hétérogène, gamma, exponentielle) furent envisagés sur données ni tronquées ni censurées. Dans

les autres parties, seule la loi normale de variance homogène fut utilisée pour générer les données, censurées à droite. L'impact de la troncature à gauche ne fut étudiée que sur données réelles.

6.3.1 MSE ou vraisemblance loi normale

Dans cette partie, nous nous sommes intéressés à la comparaison entre le réseau de neurones minimisant les moindres carrés et celui maximisant la vraisemblance de la loi normale. Bien que les programmes d'optimisation puissent être séparés pour l'espérance et la variance, il est intéressant de savoir quel est l'impact sur l'apprentissage lorsqu'on les traite ensemble.

Au cours des différents essais, la vitesse de convergence du réseau basé sur les moindres carrés fut supérieure à celle basée sur la vraisemblance. Dans le cas d'une loi exponentielle (éloignée en terme de forme) le réseau basé sur la vraisemblance divergeait et ses performances pour une loi normale homogène furent inférieurs à celui basé sur les moindres carrés.

En revanche, dans le cas d'une loi normale à variance hétérogène ou d'une loi gamma (proche en terme de forme), le réseau basé sur la vraisemblance aboutissait dans certains cas, où le volume de données pour l'apprentissage était faible et la variance suffisamment corrélée aux variables explicatives, à de meilleures performances sur la base de test. Si l'information sur la variance et la corrélation entre la variance et la moyenne sont suffisantes vis à vis de l'information directement disponible sur la moyenne dans les données, l'apprentissage joint des deux paramètres peut améliorer la qualité de prédiction de la moyenne. Ces résultats sont fortement à nuancer, les temps de calculs étant au moins un ordre de grandeur au dessus pour le modèle basé sur la vraisemblance d'une loi normale : à temps de calcul équivalent, la méthode des moindres carrés permet d'utiliser un réseau plus large et profond. Il serait intéressant d'approfondir la comparaison des deux approches.

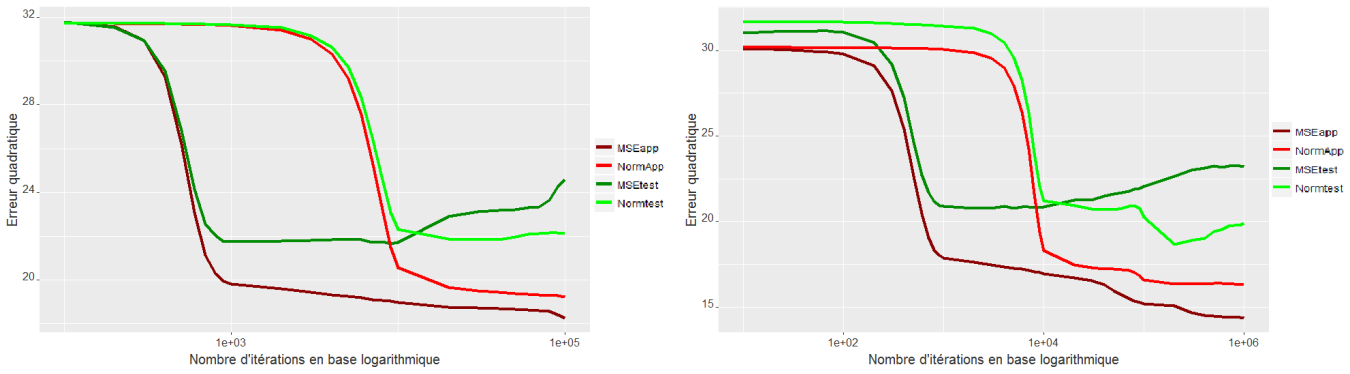


FIGURE 20: *Comparaison des erreurs du réseau basé sur les moindres carrés (couleur foncée) et du réseau basé sur la vraisemblance de la loi normale (couleur claire), sur la base d'apprentissage (rouge) et la base de test (vert), dans le cas d'une variance homogène (gauche) et hétérogène (droite)*

6.3.2 Sensibilité aux nombres de données censurées et valeurs extrêmes

Dans cette sous-partie et celle à venir, nous comparons les réseaux construits avec la correction par poids et la correction par vraisemblance générale. Nous nous intéressons aux avantages/inconvénients formulés 3 et 4 qui concernent la sensibilité des méthodes au pourcentage de données censurées et aux valeurs extrêmes.

Pour la taille de données fixée choisie, nous observons une forte dégradation des performances de la correction par poids lorsque la proportion de données censurées augmente (censure = tirage selon une loi normale indépendante des variables explicatives), les performances de la correction par vraisemblance générale restant stable, confirmant l'hypothèse 3. Notons que pour un taux de censure de l'ordre de 10%, les performances des deux méthodes sont comparables, au delà le réseau par maximum de vraisemblance est préférable : on peut supposer que pour un pourcentage de censure suffisant, la complexité introduite avec la fonction de perte de la vraisemblance générale dans la descente de gradient est compensée par la précision de la prise en compte des données de censure qu'elle permet.

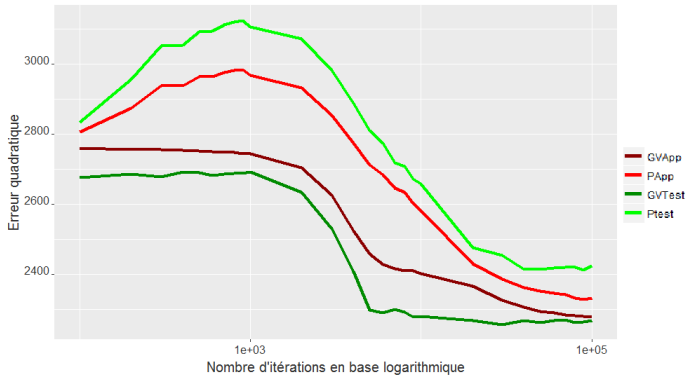
Dans la suite, afin de pouvoir comparer l'impact des autres éléments, les données seront simulées avec un taux de censure de l'ordre de 10%.

Table 2: *Impact du pourcentage de données censurées sur les performances de la correction par vraisemblance générale et la correction par poids*

% de censure	GénVraisemblance	Poids
10	2138.084	2135.931
20	2157.555	2222.058
30	2150.326	2292.356
40	2153.47	2854.287
50	2174.135	3021.318

Concernant l'étude de l'impact des valeurs extrêmes (valeur extrême simulée = moyenne 10x plus grande), les résultats sont plus difficilement interprétables. Alors que l'ajout de 1% de valeurs extrêmes dégrade les performances du réseau obtenu par poids comparativement à ceux du réseau par vraisemblance, au fur et à mesure que le pourcentage de valeurs extrêmes augmente, l'écart de performance reste identique. Nous envisageons deux façons d'interpréter ce résultat : on peut dire que les performances des deux méthodes se dégradent de manière similaire, ou encore que la frontière entre valeurs extrêmes et valeurs usuelles "disparait" lorsque le pourcentage augmente, l'erreur d'apprentissage correspondant alors au biais introduit dans les données. Il aurait fallu pour vérifier cette hypothèse, mesurer aussi l'évolution du biais dans les données simulées.

Une autre information intéressante est apparue pendant l'étude des valeurs extrêmes, qui correspond à la forme en cloche de l'apprentissage du réseau par poids : le départ de l'apprentissage commence par une phase d'accumulation de l'erreur. Cela peut s'interpréter par une attraction trop importante des poids vers les valeurs extrêmes pendant la phase où les différents neurones se spécialisent.



% de VE	GénVraisemblance	Poids
1	2257.459	2413.257
2	2414.143	2608.818
3	2500.946	2733.495
4	2559.275	2827.111
5	2516.928	2745.286
10	2814.385	2987.222

FIGURE 21: *À droite : Impact du pourcentage de valeurs extrêmes sur les performances de la correction par vraisemblance générale et la correction par poids. À gauche : comparaison des erreurs de la méthode par vraisemblance générale (couleur foncée) et de la méthode par poids (couleur claire), sur la base d'apprentissage (rouge) et la base de test (vert), pour* 1% *de valeurs extrêmes*

6.3.3 Sensibilité à la corrélation des covariables et de la censure

Il est toujours possible, en discrétisant les variables, de contourner l'hypothèse requise pour la correction par poids d'indépendance entre les covariables et la censure. Il suffit pour ce faire de calculer les pondérations sur chaque sous-population (pour chaque valeur possible de la variable) indépendamment, ce qui s'interprète comme un Kaplan-Meier conditionnel.

Cependant, cet éclatement du jeu de données a un prix : les poids sont calculés sur moins de données, et la corrélation de la censure et de la covariable considérée implique des segments où le pourcentage de censure est relativement élevé, favorisant les biais constatés dans la partie précédente. Il est donc important, pour choisir entre une hypothèse d'indépendance non respectée ou l'éclatement du jeu de données, de savoir quel est l'impact quantatif de la présence de corrélation avec la censure.

Pour cela, nous introduisons une corrélation entre la censure et l'une des variables explicatives. En jouant sur la moyenne et la variance du biais introduit, nous pouvons choisir de fixer le niveau de corrélation à la valeur souhaitée tout en conservant 10% de données censurées.

Les tableaux ci-dessous nous permettent de vérifier au passage, que la corrélation avec la censure n'a pas d'impact sur les résultats concernant la correction par vraisemblance générale. Même pour un pourcentage de censure relativement élevé, l'impact sur les performances est négligeable vis à vis de l'impact du pourcentage de censure ou celui des valeurs extrêmes calculés dans la partie précédente. Il semble alors qu'il vaille mieux ne pas respecter l'hypothèse d'indépendance plutôt que d'utiliser une pondération conditionnelle. Il aurait était intéressant de comparer directement les performances avec une correction par pondération conditionnelle.

Table 3: *Impact de la corrélation (Spearman) avec la censure pour 10% de données censurées*

Corrélation	GénVraisemblance	Poids
0	2138.084	2135.931
0.2	2143.194	2142.28
0.5	2137.354	2151.462
0.95	2136.986	2195.886

Table 4: *Impact de la corrélation (Spearman) avec la censure pour 30% de données censurées*

Corrélation	GénVraisemblance	Poids
0	2150.326	2292.356
0.5	2164.258	2286.951

6.4 Comparaison sur données réelles

Nous nous limiterons dans cette section aux données correspondant à l'incapacité. Le taux de censure étant aux alentours des 25%, on s'attend à ce que les résultats soient en faveur de la correction par vraisemblance générale.

L'âge au départ à la retraite est corrélé à la censure. Au vu de l'étude théorique et pour gagner du temps, nous admettrons qu'il est préférable de négliger cette corrélation avec la censure pour la correction par poids plutôt que de calculer des poids conditionnellement à chaque âge. Une comparaison des modèles serait néanmoins intéressante.

La mise en pratique s'avéra plus complexe que pour la simulation, avec plus de variables, plus de neurones cachés pour envisager des effets plus complexes et des temps d'apprentissage beaucoup plus longs en ordre de grandeur (10^3 fois plus d'itérations que sur les données simulées). Les graphes présentés ci-dessous représentent 5 heures de temps de calcul sur notre machine.

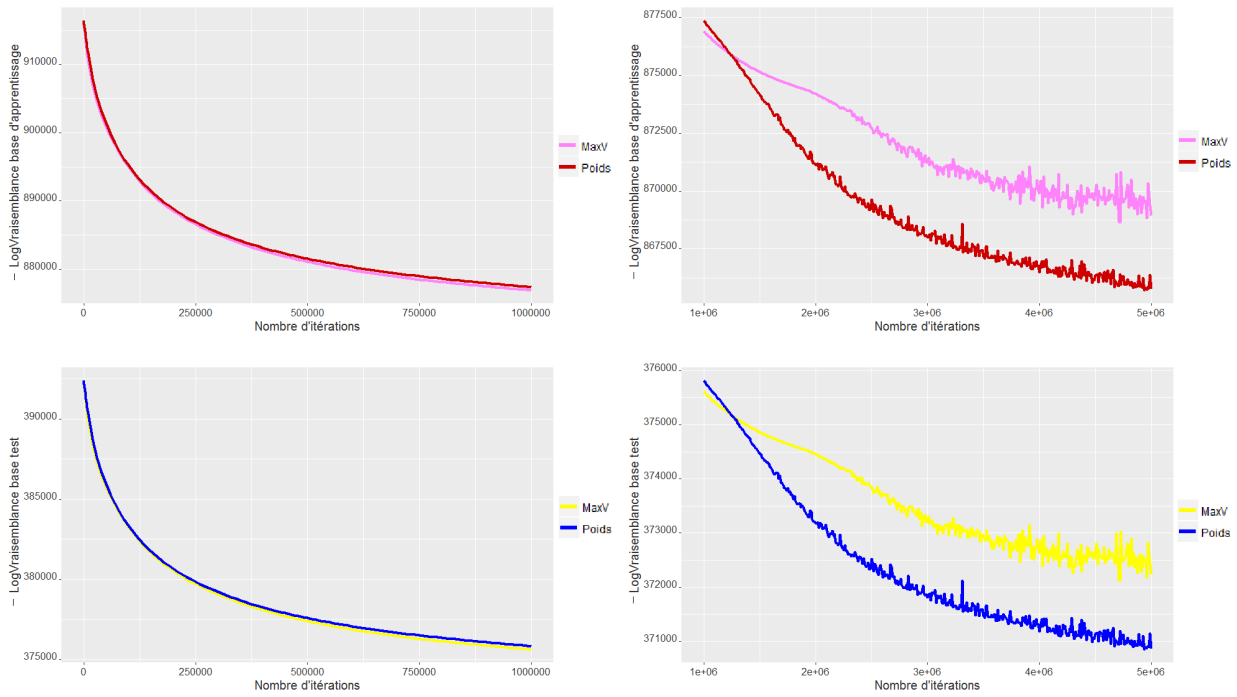


Figure 22: *Comparaison de l'évolution des erreurs pour la méthode par maximum de vraisemblance (couleur claire) et la méthode par poids (couleur foncée), sur la base d'apprentissage (haut) et la base de test (bas), pour les premières itérations (gauche) et la suite des itérations (droite)*

Contrairement à ce qui avait été envisagé, la correction par poids est plus performante que la correction par vraisemblance générale, malgré le biais introduit dans la méthode par l'absence de prise en compte de la troncature lors de la construction des poids. Parmi les pistes pour expliquer ce résultat, on peut envisager que la complexité des calculs engendrée par l'usage de la vraisemblance totale détériore l'apprentissage du modèle comparé à une méthode par poids.

7 Modèle de risque court et de risque long

7.1 Présentation du modèle

Globalement, la lecture des taux de sortie lissés (voir figure 12) nous révèle deux tendances : une tendance de court terme en mois où les taux de sortie sont décroissants et une tendance sur le long terme en années où les taux sont croissants.

La loi de Weibull de paramètre $\alpha > 0$ et $\beta > 0$ admet comme densité $f(x, \alpha, \beta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}}$, comme fonction de survie $S(x, \alpha, \beta) = e^{-\left(\frac{x}{\alpha}\right)^{\beta}}$ et donc comme taux instantané de sortie $\mu(x, \alpha, \beta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}$. Nous rappelons que les taux instantanés sont reliés au taux par période par la formule $p_t = 1 - e^{-\int_t^{t+1}\mu(t)dt}$. Un intérêt direct à cette loi est que selon la valeur de β ($\beta > 1$ ou $\beta < 1$) les taux de sortie sont croissants ou décroissants.

Une première idée serait de calibrer une loi de Weibull sur la partie incapacité et la partie invalidité pour décrire les tendances générales.

Une approche plus sophistiquée est de réfléchir en terme de risque de court terme et de long terme. Nous supposerons que ces risques ont des taux instantanés respectivement décroissant et croissant. La tendance générale de décroissance en incapacité et de croissance en invalidité serait alors le reflet de l'évolution dans le temps des proportions de ces deux types de risque.

Nous écrirons la vraisemblance d'un tel modèle comme un mélange de deux lois de Weibull dont les paramètres sont : ceux d'une loi pour le risque de court terme α_1 et β_1, ceux d'une loi pour le risque de long terme α_2 et β_2 et la proportion de risque de court terme au départ p. La fonction de survie d'un tel modèle est : $S(t, p, \alpha_1, \beta_1, \alpha_2, \beta_2) = pS(t, \alpha_1, \beta_1) + (1-p)S(t, \alpha_2, \beta_2)$ où S est la fonction de survie d'une loi de Weibull. En présence de censure et de troncature, la fonction à minimiser est alors :

$$F = -LogV = \sum_{i=1}^{n} -\delta_i ln\left[pf(Y_i, \alpha_1, \beta_1) + (1-p)f(Y_i, \alpha_2, \beta_2)\right] -$$

$$(1-\delta_i) ln\left[pS(Y_i, \alpha_1, \beta_1) + (1-p)S(Y_i, \alpha_2, \beta_2)\right] + ln\left[pS(\tau_i, \alpha_1, \beta_1) + (1-p)S(\tau_i, \alpha_2, \beta_2)\right]$$

La dérivée vis à vis des différents paramètres est calculable et notre package propose une implémentation adaptée pour maximiser la vraisemblance sur les différents paramètres simultanément. La contrainte $0 < p < 1$ est incluse au sein du modèle à l'aide de deux sous-paramètres $u, v \in \mathbb{R}$ de telle sorte que $p = \frac{e^u}{e^u + e^v}$ et $1 - p = \frac{e^v}{e^u + e^v}$. De la même manière, les contraintes sur les autres paramètres

($\alpha_1 > 0$,$\alpha_2 > 0$,$0 < \beta_1 < 1$ pour la décroissance des taux du risque de court terme et $\beta_2 > 1$ pour la croissance des taux du risque de long terme) sont directement incluses au sein du graphe.

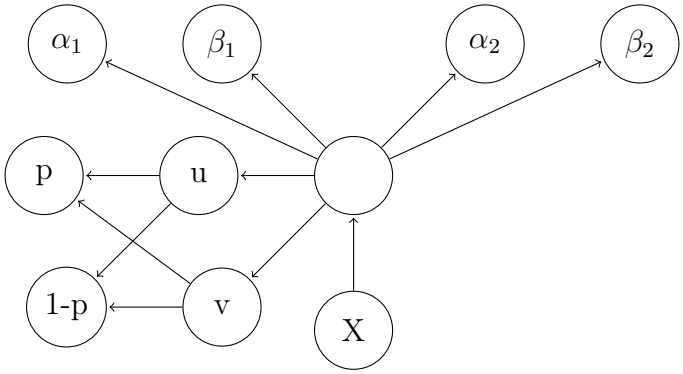


FIGURE 23: *Illustration du réseau utilisé pour le mélange de lois*

L'intérêt d'inclure une gestion de la censure est de pouvoir modéliser la loi non censurée, or nous considérons les sorties pour non prolongement en invalidité (au bout des 3 ans d'incapacité) comme des censures. Ainsi, théoriquement, la distinction entre les deux états est éliminée, la durée des arrêts de travail modélisée se comportant comme s'il n'y avait pas de séparation en incapacité et invalidité.

Cela reste néanmoins à nuancer, la distinction en incapacité et invalidité ayant au moins un impact sur le comportement des individus. Les individus prolongeant l'arrêt de travail jusqu'à la date légale de fin d'arrêt de travail, bien que marginaux, en constitue un exemple.

Cette première partie modélise la durée des arrêts de travail comme si la fin d'arrêt pour prolongement en invalidité n'existait pas. Il en découle immédiatement que pour la période correspondant à l'invalidité un tel modèle ne convient pas et nous avons alors besoin d'un second modèle sur cette partie. Nous supposons que le passage de l'incapacité à l'invalidité n'a pas d'effet sur la durée des arrêts de travail mais que les arrêts associés à un risque de court terme ont une probabilité plus faible d'être prolongés en invalidité que les arrêts associés en invalidité. Ce second modèle adapté à l'invalidité s'obtient alors à partir du modèle précédent en recalibrant uniquement la proportion de risque de court terme p sur les données correspondant à l'invalidité.

La modélisation de la durée des arrêts de travail s'effectue avec une première partie comme s'il n'y avait pas de limite de durée pour l'incapacité, avec une décomposition en terme de risque de court terme et risque de long terme par un mélange de lois. Une seconde partie prend en compte l'impact de la transition entre l'incapacité et l'invalidité, la durée des risques de court et long terme n'étant pas affectée mais leur proportion évoluant, un risque de court terme ayant plus de chance de ne pas être prolongé qu'un risque de long terme.

7.2 Utilisation des spécificités du modèle

Nous rappelons que les différents paramètres du modèle dépendent des caractéristiques des individus au début de l'arrêt et de la nature de l'arrêt. L'étude à l'échelle agrégée du modèle dépend de la répartition des variables explicatives et donc des données utilisées. Nous noterons p^1 la proportion de risque de court terme avant le passage en invalidité et p^2 la proportion post passage en invalidité.

Les quantités $100 * \frac{\sum_i p^{1,i} S(t,\alpha_1^i,\beta_1^i)}{\sum_i p^{1,i} S(0,\alpha_1^i,\beta_1^i)} = 100 * \frac{\sum_i p^{1,i} S(t,\alpha_1^i,\beta_1^i)}{\sum_i p^{1,i}}$ et $100 * \frac{\sum_i (1-p^{1,i}) S(t,\alpha_2^i,\beta_2^i)}{\sum_i (1-p^{1,i})}$ sont les pourcentages théoriques restant dans la base de risque de court et de long terme à l'instant t avant impact du passage en invalidité.

Ces informations sont intéressantes en particulier à l'instant $t = 3$ ans délimitant la fin d'incapacité : les proportions permettent de vérifier si la majorité des arrêts pour le risque de court terme se sont terminés avant le passage en invalidité et quelles proportions d'arrêts pour le risque de long terme finissent prématurément. On pourrait même envisager, si un modèle parfaitement calibré était à disposition, de réfléchir à une définition d'une durée limite d'incapacité telle que $x\%$ des arrêts pour le risque de court terme se soient terminés avant.

Une information assez similaire est la proportion d'arrêts pour le risque de court terme vis à vis des arrêts restant à l'instant t avant impact de l'invalidité $p_s^1 = 100 * \frac{\sum_i p^{1,i} S(t,\alpha_1^i,\beta_1^i)}{\sum_i p^{1,i} S(t,\alpha_1^i,\beta_1^i) + (1-p^{1,i}) S(t,\alpha_2^i,\beta_2^i)}$.

Si à un instant t la proportion des arrêts du risque de long terme dépasse les 50% ou un autre pourcentage choisi des arrêts restants, on peut se dire que nous assistons à la transition à l'échelle agrégée d'un risque court à un risque long et qu'il s'agit donc d'une bonne démarcation pour passer de l'incapacité à l'invalidité.

De même il est possible de définir pour la deuxième partie, post impact de la démarcation entre incapacité et invalidité, la proportion d'arrêt de risque court $p_s^2 = 100 * \frac{\sum_i p^{2,i} S(t,\alpha_1^i,\beta_1^i)}{\sum_i p^{2,i} S(t,\alpha_1^i,\beta_1^i) + (1-p^{2,i}) S(t,\alpha_2^i,\beta_2^i)}$ et de comparer ces deux proportions. L'interprétation est assez immédiate et permet de comprendre comment évoluent les proportions des deux risques étant donné qu'un arrêt de risque court a plus de chance de ne pas être prolongé qu'un arrêt de risque long.

L'état d'invalidité se focalisant sur le risque de long terme, nous supposerons que les arrêts de long terme ne sont pas affectés par la transition et que la variation dans la répartition des risques n'est due qu'à l'absence de prolongation en invalidité d'une partie des risques de court terme. Cette hypothèse permet de calculer à l'échelle individuelle une probabilité d'être censuré p_c^i lors du passage de l'incapacité à l'invalidité et d'en déduire un nombre moyen de censure théorique $\sum_i p_c^i$. Un moyen de tester la qualité du modèle est alors d'utiliser cette espérance et la variance associée $\sum_i p_c^i (1 - p_c^i)$ pour vérifier que le nombre de censure lors de la limite des 3 ans en incapacité est dans l'intervalle de confiance.

Exemple pour un arrêt : supposons que $p_s^1 = 90\%$ et $p_s^2 = 80\%$. Au départ, cet arrêt compte donc comme 0.9 arrêt de court terme et 0.1 arrêt de long terme. Après impact de l'invalidité, le nombre d'arrêt de long terme 0.1 n'est pas modifié mais la proportion est alors de 80% d'arrêt de court terme.

Pour respecter la proportion, le nombre d'arrêt de court terme chute alors a $\frac{0.8*0.1}{1-0.8} = 0.4$, pour un total d'arrêt de 0.5. La probabilité pour cet arrêt d'être censuré est alors $100 * \frac{0.5}{1} = 50\%$.

Afin de vérifier la qualité du modèle, et de l'utiliser avec les restrictions imposées par la législation actuelle, il est nécessaire de le convertir en table de taux de sortie. Cela se fait facilement, avec la formule pour la fonction de survie disponible et les performances peuvent alors être comparées aux tables construites par les autres modèles grâce au cadre défini dans la partie I.

À noter que les résultats de la comparaison des tables de taux sont légèrement en défaveur du modèle, la mesure des performances ne correspondant pas au programme de maximisation contrairement aux estimateurs classiques.

Si la législation le permettait, il serait préférable pour la tarification et le provisionnement de modéliser directement des durées. L'utilisation des tables de taux pour les durées passées en cas de sortie nécessite des hypothèses sur la forme de la répartition (répartition uniforme en cas de sortie dans le mois par exemple) qui sont en réalité déjà faites dans notre modèle.

Un dernier avantage de ce modèle est de faciliter la formulation d'hypothèses et leur impact sur le modèle si un changement sur la durée maximum d'incapacité venait à avoir lieu. L'hypothèse la plus simple serait que les proportions au départ des deux parties du modèle p_1 et p_2 ne varient pas, déplaçant simplement la frontière d'utilisation des deux parties. On pourrait aussi supposer par exemple que la variation de proportions à la démarcation des deux états $p_s^1 - p_s^2$ est constante.

Conclusion

Ce travail nous a permis de revisiter la construction de tables d'expériences pour le maintien en arrêt de travail. Les modèles habituellement utilisés, la détermination des taux bruts avec l'estimateur de Kaplan-Meier, les méthodes de lissages des taux bruts par splines cubiques ou avec le lissage de Whittaker-Henderson et l'ajout de covariables avec le modèle de Cox furent replacés dans le cadre général des méthodes d'apprentissage.

La sélection du degré de lissage idéal pour passer des taux bruts très irréguliers aux taux lissés fut effectuée en maximisant la vraisemblance des données sur un échantillon de validation plutôt que par une méthode visuelle moins rigoureuse.

La vérification de l'adéquation des paramètres des modèles fut remplacée par un calibrage des paramètres passant par une mesure des performances sur une base de validation. Cette méthode permit une détermination plus fine des coefficients de pénalisation des lissages, et fut comparable à un test de proportionnalité pour vérifier la pertinence du modèle de Cox. Comme on peut s'y attendre, l'élimination des variables non significatives n'a eu que peu d'impact sur les performances du modèle. Le phénomène inverse n'a pu être évalué, à savoir comment varient les performances si l'on inclut à tort une variable pour sa significativité.

Dans la seconde partie de notre travail, nous avons posé les bases nécessaires à la réalisation d'un modèle de durée à mélange de lois, censé retranscrire la distinction entre l'existence d'un risque court et d'un risque long que suppose le découpage en incapacité et invalidité.

Nous avons alors présenté l'intérêt des réseaux de neurones en tant qu'outils de calcul pour la maximisation par rétropropagation ou pour les modèles plus complexes qu'ils permettent de réaliser en espérant voir le développement de leur utilisation en actuariat

De même, nous espérons que la présentation des méthodes de corrections de modèles pour la gestion de la censure et de la troncature incite à d'avantages d'innovation dans cette direction en assurance de personnes. La comparaison des deux méthodes de correction aura permis, bien que limitée au cadre de l'étude, d'en dégager des points forts et des inconvénients et de présenter les détails d'une implémentation simple de réseaux de neurones.

Nous avons finalement proposé un modèle de mélange de lois nous paraissant prometteur pour la modélisation du maintien en arrêt de travail, ainsi que des détails sur son implémentation concrète et son utilisation. Le calcul des proportions de risques courts et de risques longs, avant et après la transition en invalidité, permettant d'éclairer la législation.

La limite principale de cette étude est que nous n'avons pas eu le temps de réaliser le dernier modèle proposé, sa mise en place par le biais d'un perceptron multi-couches nécessitant d'après nos estimations quelques mois de travail supplémentaires. Alors que les paramètres du modèle fourniraient des premiers résultats pour mieux comprendre la séparation en incapacité et invalidité, la pertinence d'un tel modèle en terme de performance serait aussi vérifiable grâce au cadre posé et aux résultats obtenus dans la première partie.

Références

Adler, D., Gläser, C., Nenadic, O., Oehlschlägel, J., and Zucchini, W. (2014). *ff : memory-efficient storage of large data on disk and fast access functions*. R package version 2.2-13.

Aubin, I. and Rolland, A. (2010). Lignes directrices de la construction des lois de maintien en incapacité et invalidité. *Institut des Actuaires*.

Bagui, H. (2013). Refonte des lois de maintien en incapacité temporaire de travail. *Université Claude Bernard*.

Bellina, R. (2014). Méthodes d'apprentissage appliquées à la tarification non-vie. *Université Claude Bernard*.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference understanding aic and bic in model selection. *Sociological methods and research, Vol. 33, No. 2*.

de Freitas, J. F. G. (2000). Bayesian methods for neural networks. *Trinity College, University of Cambridge and Cambridge University Engineering Department*.

Dowle, M. and Srinivasan, A. (2017). *data.table : Extension of 'data.frame'*. R package version 1.10.4.

Dupré, X. (2004). Contributions à la reconnaissance de l'écriture cursive à l'aide de modèles de markov cachés. *Université René Descartes*.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York. ISBN 978-1-4614-6867-7.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo : Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71 :1054–1063.

Fong, J. H., Shao, A. W., and Sherris, M. (2013). Multi-state actuarial models of functional disability. *The North American Actuarial Journal Volume 19, Issue 1*.

Freedman, D. A. (2008). Greenwood's formula.

Frees, E. W., Meyers, G., and Cummings, A. D. (2012). Insurance ratemaking and a gini index. *Journal of Risk and Insurance*.

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert*, 6(4) :46–51.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition*. Springer-Verlag New York.

Lopez, O. (2007). Reduction de dimension en presence de donnees censurees. *ENSAE Paristech*.

Lopez, O. (2013). Duration models.

Lopez, O., Milhaud, X., and Therond, P. (2015). Tree-based censored regression with applications to insurance.

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(11) :1–23.

Planchet, F. (2016). Modèles de durée méthodes de lissages et d'ajustement.

Ridgeway, G. (2007). Generalized boosted models a guide to the gbm package.

Rodriguez, G. (2010). Chapter 7 survival models.

Saint-Pierre, P. (2015). Introduction à l'analyse des durées de survie. *Université Pierre et Marie Curie*.

Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data : Extending the Cox Model*. Springer, New York.

Weng, Y. (2007). Baseline survival function estimators under proportional hazards assumption. *National University of Kaohsiung*.

Wood, S. (2006). *Generalized Additive Models : an introduction with R*. Chapman and Hall/CRC.

Wood, S. (2016). Package mgcv cran : Mixed gam computation vehicle with gcv aic reml smoothness estimation.

Table des figures

Liste des tableaux

8 Annexes

8.A Critères d'une fonction de perte

Nous avons vu que la question de l'apprentissage supervisé peut s'écrire comme la recherche d'une fonction minimisant un certains critère $\mathcal{C}$.

Il est commun de définir à chaque réalisation quelle est la perte générée par notre estimation, ce qui s'écrit par la détermination d'une fonction, dite fonction de perte $\mathcal{L} : (y, f(x)) \to \mathcal{L}(y, f(x))$. Il est commun de sélectionner $\mathcal{L}$, vérifiant certaines propriétés :

— pas de gain possible : $\mathcal{L}$ à valeurs dans $\mathbb{R}^+$ (critère non respecté par la mesure des écarts en tarification ou provisionnement)

— une perte nulle pour une prédiction parfaite : $\mathcal{L}(y, y) = 0$

On détermine ensuite à partir de cette fonction un critère à minimiser, correspondant à une mesure du risque encouru. Dans notre cas, nous nous intéressons au risque espéré, c'est à dire : $\mathcal{C}(Y, f(X)) = \mathrm{E}_{Y,X}[\mathcal{L}(Y, f(X))]$

8.B Note sur la mesure des performances de méthodes d'apprentissage en assurance à l'aide d'un indice Gini

Cette note explique de manière succincte le critère présenté dans (Frees et al., 2012) en vue de comparer des méthodes de prédictions de la prime pure en population générale. Après un rappel sur la courbe de Lorenz et l'indice Gini, nous verrons la courbe de Lorenz ordonnée utilisée et l'indice Gini associé ainsi que ses propriétés d'interprétation et enfin comment l'utiliser.

8.B .a Rappels sur la courbe de Lorenz et l'indice Gini

La courbe de Lorenz est un outil qui fut développé en 1905 par Max Otto Lorenz, utilisé à l'origine pour étudier les distributions de revenu : il s'agissait alors du graphe de la fonction de répartition des revenus en fonction de la proportion de la population considérée. Les différents points (x,y) de la courbe permettaient de déterminer quel proportion y de richesse les x% plus pauvres détenaient. La diagonale correspondait alors à la situation où les richesses sont parfaitement réparties entre les individus, tout écart à cette diagonale représentant une situation d'inégalité. Corrado Gini en 1912 eut l'idée de prendre comme mesure d'inégalité le double de l'aire entre la courbe tracée et la ligne d'égalité (la diagonale), quantité nommée par la suite indice de Gini.

L'usage de la courbe et de l'indice s'est généralisé à d'autres domaines, pour son utilité à comparer des distributions complexes. La question de l'efficacité des méthodes d'apprentissage en assurance, relevant des montants de remboursement dont la distribution est asymétrique (vers la droite) avec un pic en zéro, semble alors adaptée à l'usage d'un tel indice contrairement à des mesures usuelles en apprentissage tel que l'erreur absolue ou l'erreur quadratique.

Une première idée afin de comparer les méthodes pourrait être de superposer les courbes de Lorenz des montants observés et des montants prédits. Un tel graphe n'aurait cependant aucune interprétation, puisque les deux courbes relèvent de classements sur les individus différents. On utilise alors un critère permettant d'établir un classement commun aux deux distributions, en vue de calculer un indice gini. Nous verrons alors comment l'indice gini calculé relève la vulnérabilité d'une méthode par rapport à une autre dans une optique de sélection adverse et permet alors de sélectionner la "meilleure" méthode. Similairement aux autres critères, l'indice obtenu est à évaluer sur une base de test différente de la base d'apprentissage.

8.B .b Présentation et interprétation

Nous nous intéressons à la question de la tarification des contrats des N individus disponibles dont les sinistres entraînent un montant à charge à l'enterprise (une perte) y_i. Nous ne considérerons pas les divers chargements à rajouter, ainsi le tarif proposé découle directement de la prime pure, que l'on cherche à déterminer à l'aide des méthodes d'apprentissage. À partir des caractéristiques $X_1, ..., X_N$ des individus, nous cherchons à comparer les tarifs prédits d'une méthode d'apprentissage principale $P(X_i) = P_i$ à ceux d'une méthode concurrente S_i. Nous supposerons par la suite que les budgets sont normalisés : $\mathrm{E}[L] = \mathrm{E}[P] = \mathrm{E}[S] = 1$.

La courbe de Lorenz ordonnée s'obtient à partir d'un critère R permettant de classer les individus : il s'agit du graphe empirique de la distribution associée des montants à charge $F_L(s) = \mathrm{E}[L|R \leq s]$ en fonction de la distribution des tarifs (les recettes) $F_P(s) = \mathrm{E}[P|R \leq s]$, qui s'écrivent (en normalisant à un) :

$$\hat{F}_L(s) = \frac{\sum_{i=1}^N y_i \mathbb{1}_{R_i \leq s}}{\sum_{i=1}^N y_i} \text{ et } \hat{F}_P(s) = \frac{\sum_{i=1}^N P_i \mathbb{1}_{R_i \leq s}}{\sum_{i=1}^N P_i}$$

Il est alors possible de montrer que l'indice de Gini calculé $\widehat{Gini}$ correspondant à l'aire sous la courbe est, sous certaines conditions de régularité, un estimateur sans biais avec une distribution asymptotique normale de l'indice de Gini théorique $Gini = 2\int_0^\infty (F_p(s) - F_L(s))dF_P(s)$.

Pour $F_P(s) - F_L(s) \geq 0$, nous sommes sur un segment de population profitable à l'entreprise puisque les recettes sont supérieures aux dépenses. L'indice de Gini $\widehat{Gini} \approx \frac{1}{N}\sum_{i=1}^N (\hat{F}_P(R_i) - \hat{F}_L(R_i))$ peut alors s'interpréter comme une moyenne des profits obtenables selon les stratégies (les sous-populations) définies par la connaissance du critère R. Cependant, nous rappelons que nous avons supposé que nous étions dans une situation d'équilibre entre les dépenses et les recettes : une sous-population où l'entreprise dégage un profit est une population sur-tarifée, qui implique l'existence d'une population sous-tarifiée pour laquelle l'entreprise présente un déficit. L'indice calculé peut ainsi s'interpréter comme une vulnérabilité de l'entreprise ne séparant pas efficacement les populations à risques sur les segments définis par le critère R.

Le critère de classement utilisé correspond au tarif relatif de la méthode concurrente vis à vis de la méthode principale $R_i = \frac{S_i}{P_i}$ appelé relativité. Avec le raisonnement précédent, la connaissance du tarif concurrentiel permet de dégager des segments profitables, l'indice de Gini se traduisant comme

une vulnérabilité de la méthode principale à la méthode concurrentielle.

Même si la concurrence ne connaît pas le tarif de la méthode principale, l'indice de gini est interprétable comme une vulnérabilité de la méthode par sélection adverse. Nous avons sous certaines approximations ($\hat{F}_P(R)$ approximable par $\hat{F}_R$) $\widehat{Gini} \approx \frac{2}{N}\widehat{Cov}(PP, rang(R))$ où PP est la charge rapportée à la prime pure (y/P) : un indice de gini élevé correspond à une covariance importante entre la charge rapportée et le rang du critère.

Dans le cadre d'un indice de gini élevé, en supposant que la concurrence emploie comme méthode de tarification la méthode secondaire S à budget équivalent, les individus pour lesquels le tarif de la concurrence est préférentiel (R faible) correspond à des individus de charge rapportée faible, une sous-population profitable. À l'inverse, les individus pour lesquels le tarif principal est préférentiel (R élevé) correspond à des individus non rentable de charge rapportée élevée : la méthode principale n'effectuant pas une distinction correcte sur la sous-population concernée se retrouve par le jeu de la sélection adverse déficitaire, ne conservant que le "mauvais risque" le "bon risque" étant capté par la concurrence.

Que la relativité soit connue ou non de la concurrence, l'indice de Gini est une mesure de la vulnérabilité de la méthode principale à la méthode concurrente d'un point de vue de la sélection adverse.

8.B .c Utilisation pour la comparaison de plusieurs méthodes

Un moyen de sélectionner la méthode de tarification la moins "vulnérable" à la concurrence parmi plusieurs disponibles est de calculer les différents indices de Gini, en considérant chaque méthode comme la principale et la secondaire à tour de rôle. La méthode la moins exposée à la concurrence est celle pour laquelle le pire choix de la concurrence (relativement à la méthode) aboutit à la meilleure situation : on choisit la méthode dont le maximum des indices est le plus bas.

Une confirmation de la qualité de la méthode que l'on observe parfois, est que celle-ci est celle causant l'indice de gini le plus élevé chez les autres méthodes.

Ce critère de sélection ne permet pas toujours d'aboutir à un classement des méthodes, que ce soit pour la significativité de la différence des indices (voir plus bas) ou si l'on obtient des indices égaux. Il est possible en effet qu'aucune méthode ne soit pleinement satisfaisante, dans le sens où celles-ci sont incomplètes : chacune arrivant à distinguer une sous-population mal tarifée par l'autre méthode.

L'avantage de cette mesure des résultats vis à vis d'autres méthodes (erreur quadratique ou absolue) est la possibilité de calculer des intervalles de confiance (voir (Frees et al., 2012)) en vue de vérifier la significativité de l'indice sans avoir à les estimer empiriquement, permettant d'éviter l'usage de méthodes plus coûteuse qu'une simple séparation apprentissage-test pour comparer les méthodes (telle que la validation croisée).

Le package R cplm implémente les différents calculs à effectuer, notamment celui des intervalles de confiance. Les calculs restent simples et peuvent être facilement réécrits dans d'autres languages.

Quelques vérifications supplémentaires sont cependant nécessaires pour compléter la comparaison des méthodes : à l'instar de la courbe ROC et du critère AUC, l'indice ne change pas selon une transformation strictement croissante de l'une des variables étudiées. Même si celles-ci ne sont plus utilisées directement pour la comparaison de modèle, le calcul des erreurs quadratique et absolue permet de fournir des informations supplémentaires sur la répartition des prédictions. Enfin, de nombreux résultats théoriques et l'interprétation en terme de sélection adverse repose sur le concept de budget normalisé, la bonne utilisation du critère impose de vérifier au préalable que cette hypothèse soit valide, c'est à dire $\sum_{i=1}^{N} y_i \approx \sum_{i=1}^{N} P_i \approx \sum_{i=1}^{N} S_i$.

8.C Courbes de taux pour le lissage de Whittaker-Henderson

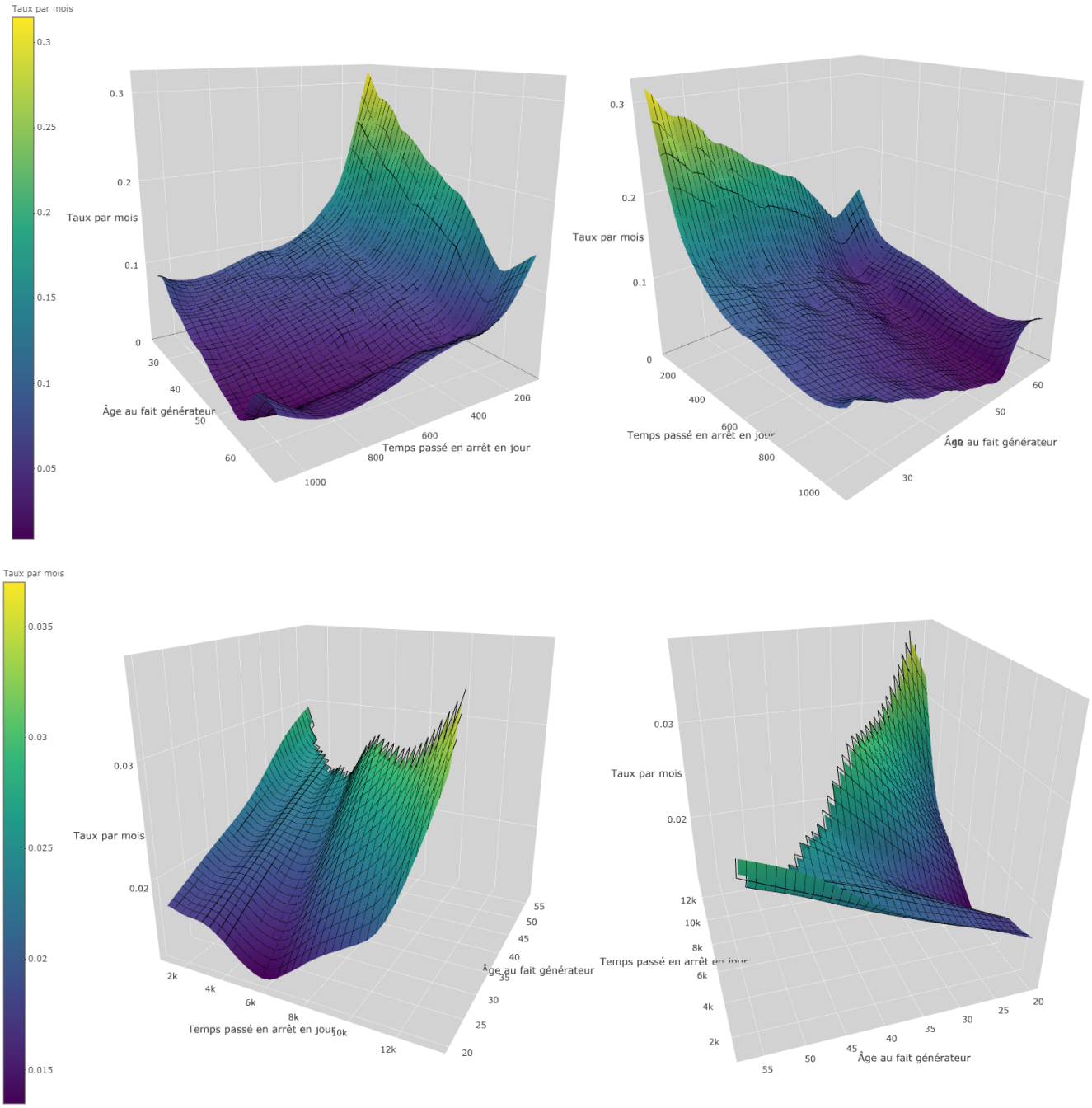


Figure 24: *Graphes des taux lissés par Whittaker-Henderson pour la partie correspondant à l'incapacité (haut) et la partie correspondant à l'invalidité (bas) sous deux angles différents*

8.D Critère AIC et BIC : présentation et comparaison

Ce n'est que tardivement que la sélection de variables fut déterminée par sélection du modèle d'erreur minimum sur un échantillon de validation, pour l'homogénéisation du calibrage. Avant cela, la méthode reposait sur un critère qu'il fallut sélectionner, entre AIC et BIC. Cette sélection fut effectuée à l'aide de la réflexion portée dans (Burnham and Anderson, 2004), brièvement présentée dans cette annexe.

Nous supposons qu'il existe une véritable distribution à densité f de nos données, dont s'écarte la densité approximée de notre modèle $f(.,\theta)$. Une mesure de la dissimilarité entre deux distributions est la distance de Kullback-Leibler, prenant la forme :

$$D_{KL}(f\|g) = \int_{\mathbb{R}} f(x)log(\frac{f(x)}{g(x)})dx = \underbrace{\int_{\mathbb{R}} f(x)log(f(x))dx}_{\text{ne dépend pas de g}} - \int_{\mathbb{R}} f(x)log(g(x))dx$$

Le choix entre deux modèles peut donc s'effectuer uniquement sur le second membre $\mathrm{E}_X[log(g(X))]$.

Etant donné que les modèles passent par une estimation de la fonction, Akaike montra qu'un critère rigoureux de sélection de modèle devait se baser sur :

$\mathrm{E}_{\mathcal{A}}\mathrm{E}_X[log(\hat{f}(X)]$ où $\mathcal{A}$ base d'apprentissage et θ estimé par maximum de vraisemblance

Bien que la vraisemblance maximisée sur la base d'apprentissage soit un estimateur biaisé de ce critère, Akaike trouva que ce biais est proportionnel au nombre de paramètres du modèle, et proposa alors comme critère de comparaison nommé AIC (Akaike information criteria) :

$AIC = -2log(Vraisemblance\ maximisee \mid donnees) + 2K$ où K nombre de paramètres du modèle

Le critère BIC (Bayesian Information Criterion) fut par la suite dérivé du critère AIC, en vue de le corriger : $BIC = -2log(Vraisemblance) + Klog(n)$ où n est la taille de l'échantillon. Cette correction permet de s'assurer de la convergence du modèle : là où le modèle sélectionné par le critère AIC dépend de la taille de l'échantillon, et favorise les modèles trop complexes (n'augmentant pas la distance à la distribution significativement), le critère BIC converge, vers un modèle concentrant une partie de l'information. La convergence ne saurait être un argument pour sélectionner le critère BIC, la notion de taille d'échantillon suffisamment grande étant floue, celui-ci pouvant s'éloigner significativement de son modèle cible quelle que soit le nombre de données.

Un autre argument pour sélectionner le critère BIC fut son lien avec une approche bayésienne du problème. En considérant que chacun des modèles envisagés à la même probabilité à priori, pour chaque modèle g_i la probabilité à postériori peut s'écrire :

$$\mathbb{P}(g_i \mid donnees) = \frac{e^{-\frac{1}{2}\Delta BIC_i}}{\sum_{r=1}^{R} e^{-\frac{1}{2}\Delta BIC_r}}$$ où R est le nombre de modèles envisagés

Mais cet argument est invalide, étant donné la ressemblance entre les deux critères, on peut corriger les lois à priori de telle sorte que la probabilité à postériori s'écrit cette fois-ci à partir du critère AIC. Cette écriture est même parfois préférée, étant donné que la loi à priori utilisée pénalise les modèles les plus complexes.

Le choix dépend finalement de la pertinence de la cible des deux critères, et donc de la nature de la véritable distribution : pour une distribution approximable en quelques paramètres le critère BIC permet de ne conserver que les éléments clés pour la décrire, alors que pour une distribution complexe

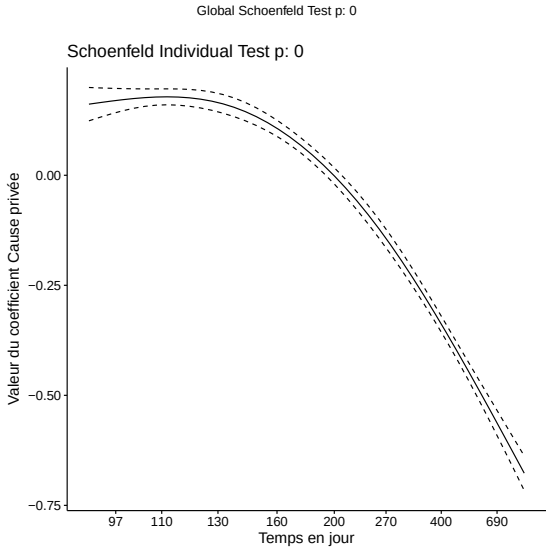
le critère AIC permet de conserver un maximum de l'information pour l'appréhender. Il a cependant été démontré, sous peine que le véritable modèle ne soit pas parmi ceux de la sélection, que le critère AIC minimisant la distance K-L permet d'aboutir au modèle d'erreur quadratique minimale. En vue de nos objectifs, le critère AIC fut celui conservé.

8.E Informations sur le modèle de Cox obtenu

Coefficient	Valeur	Variance	Valeur test	p-valeur test
		Incapacité		
Sexe M	−0.013	0.012	−1.1	0.27
Cause privée	−0.054	$5.6e-3$	−9.7	$< 2e-16$
CSP Etam	−0.012	0.015	−0.79	0.43
CSP Ouvrier	−0.066	0.013	−5.0	$4.8e-7$
		Invalidité		
Sexe M	0.30	0.096	3.1	$1.8e-3$
Cause privée	0.77	0.11	6.8	$1.2e-11$
CSP Etam	0.098	0.11	0.89	0.37
CSP Ouvrier	0.25	0.094	2.7	$7.6e-3$

Figure 25: *Modèle de Cox obtenu : coefficient, variance, test de significativité et p-valeur pour l'incapacité et l'invalidité*

Par exemple, les ouvriers en invalidité ont des taux de sortie $\exp(0.25) \approx 1.29$ fois plus élevé que ceux des cadres, la p-valeur du test étant inférieur à 0.01, on peut rejeter l'hypothèse de nullité du coefficient au seuil de 1% et considéré son importance comme significative.



Coefficient	ρ	χ_2	$p-valeur$
		Incapacité	
Sexe M	$-7e-3$	7.8	$5.3e-3$
Cause privée	−0.11	1850	$< e-10$
CSP Etam	$1.0e-3$	0.17	0.68
CSP Ouvrier	0.013	26	$2.7e-7$
		Invalidité	
Sexe M	0.019	1.1	0.29
Cause privée	$-1.7e-3$	$9.5e-3$	0.92
CSP Etam	−0.016	0.84	0.36
CSP Ouvrier	$-2.2e-3$	0.015	0.23

Figure 26: *Tableau résumant les différents tests de proportionnalité réalisés (droite) et graphique d'illustration de la moyenne des résidus de Schoenfeld et leur variance en fonction du temps pour une cause privée d'arrêt (gauche)*

Par exemple, comme on peut le confirmer graphiquement, la p-valeur du test des résidus de Schoenfeld concernant la nature de la cause en incapacité est inférieur à 0.01, on peut donc rejeter l'hypothèse de proportionnalité (absence de variation dans le temps) du coefficient.

8.F Graphes du modèle de Cox

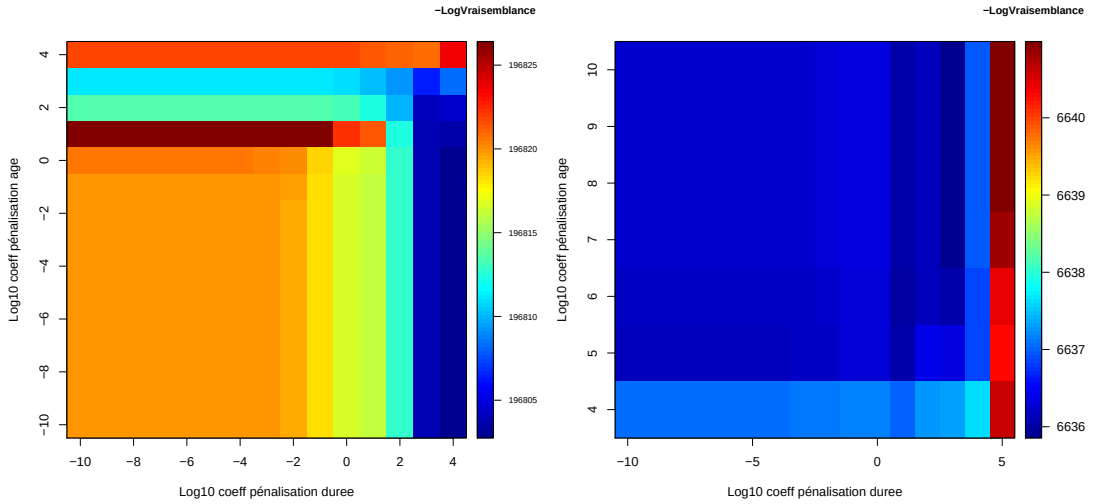


Figure 27: *Opposé de la logVraisemblance sur l'échantillon de validation pour le calibrage du lissage par splines du modèle de Cox, pour la partie concernant l'incapacité (gauche) et l'invalidité (droite)*

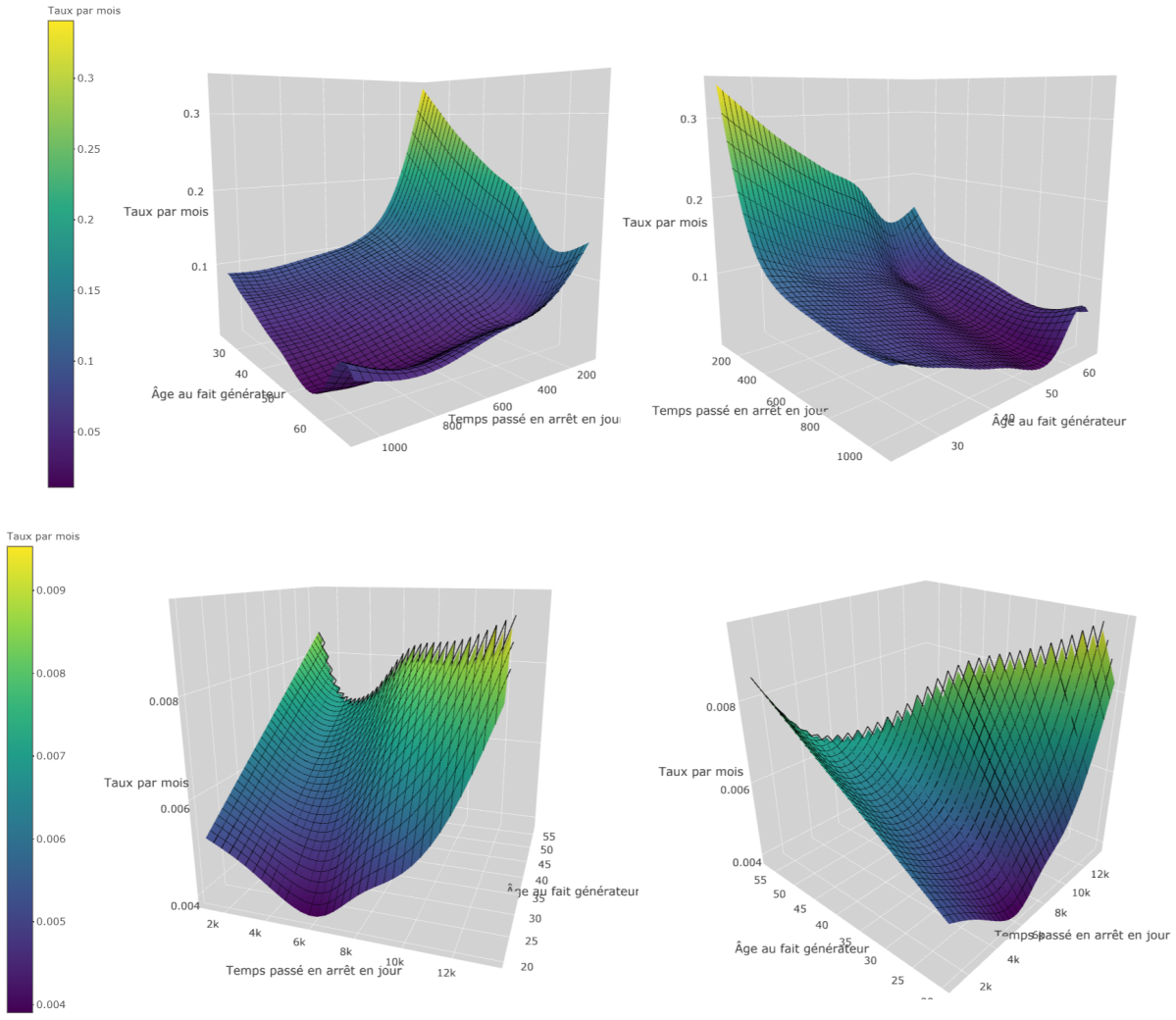


Figure 28: *Graphes des taux lissés par splines cubiques du modèle de Cox pour la partie correspondant à l'incapacité (haut) et la partie correspondant à l'invalidité (bas) sous deux angles différents*

Note de synthèse

Modélisation du maintien en arrêt de travail

Tristan JUDD

14 Février 2018

Les contrats collectifs traitant de l'arrêt de travail temporaire et de l'arrêt de travail permanent (invalidité) constituent une part significative, de plus de deux milliards d'euros, du passif de BTP-Prevoyance. Des gains mêmes de quelques pourcents sur les montants de provisionnement représentent des enjeux importants pour l'optimisation des fonds propres de l'entreprise.

Cette tarification est en général réalisée à l'aide d'un modèle fréquence-durée, et nous nous sommes concentrés sur la modélisation du maintien en arrêt de travail pour le calcul du provisionnement, qui depuis la loi Evin, se fait à partir de tables de survie ou des tables de taux permettant d'estimer le temps moyen de l'arrêt en incapacité ou invalidité. Bien que des tables officielles soient mises à disposition, il est recommandé de réaliser des tables d'expérience, afin de prendre en compte le plus fidèlement possible les caractéristiques spécifiques du portefeuille de l'assureur.

Les durées étudiées s'étalent sur plusieurs années voire décennies, et il n'est pas envisageable de bénéficier d'observations couvrant l'ensemble du phénomène. La construction des tables s'est alors limitée par le passé à l'usage de méthodes spécifiquement adaptées. Il est désormais possible, par le biais de correctifs, d'adapter une grande majorité des méthodes d'apprentissage pour couvrir ces spécificités.

Cette étude présente les moyens disponibles pour l'implémentation de ces méthodes, en particulier par l'usage des réseaux de neurones, qui sont encore à l'heure actuelle inutilisés dans la majorité des travaux de recherche en actuariat. Ces nouvelles méthodes permettent des projets de modélisations plus ambitieux, non plus centrés sur les états d'incapacité et d'invalidité, mais cherchant à mesurer directement le phénomène de l'arrêt de travail. Un tel modèle ferait implicitement la distinction entre les deux états, classant les arrêts en deux catégories, relevant d'un risque court et d'un risque long.

La première partie de l'étude s'est concentrée sur la réalisation et la mise au goût du jour des méthodes usuelles. Elle nous aura permis de les replacer au sein d'un cadre théorique et méthodologique permettant la comparaison avec des modèles de durée, une fois ceux-ci convertis en table de taux.

Les données construites regroupèrent les informations nécessaires : les dates de début et fin d'observation, la cause de fin d'observation (fin d'arrêt ou censure), les variables nécessaires à la construction des tables (l'âge au fait générateur et le temps passé dans l'état) et des variables explicatives supplémentaires. Une analyse statistique descriptive rapide mit en avant les spécificités du portefeuille, en particulier la proportion plus importante d'hommes et d'ouvriers qu'en population générale.

Après une présentation du cadre des méthodes d'apprentissage et de la méthodologie de comparaison, les taux bruts furent calculés à l'aide de l'estimateur de Kaplan-Meier. Ces taux furent ensuite lissés, avec deux méthodes distinctes, que sont le lissage par splines cubiques et le lissage de Whittaker-Henderson. Habituellement, les coefficients de lissage, réglant le degré de lissage des courbes, sont déterminés visuellement et vérifiés à partir de tests. Nous avons préféré une sélection plus précise des coefficients, en maximisant les performances sur un échantillon de vérification, mesurées comme la différence entre les taux observés et les taux prédits. Les méthodes de lissage utilisées sont assez proches théoriquement ainsi que sur la forme des courbes obtenues, ce qui nous a permis de fixer un ordre de grandeur sur la comparaison des performances. Une méthode est jugée significativement meilleure si ces performances se distinguent de celles des deux lissages.

La prise en compte des variables supplémentaires fut effectuée à l'aide d'une méthode classique : le modèle de Cox. Celui-ci repose sur l'hypothèse de proportionnalité des variables ajoutées, c'est à dire que leur impact ne varie pas au cours du temps. Le choix de réaliser ou non le modèle dépend souvent de la vérification de cette hypothèse de proportionnalité, nous avons ici préféré vérifier l'impact de la validité de cette hypothèse sur les performances. La sélection de variable ne fut, de même, pas effectuée par un test de leur significativité, mais par leur impact en terme de performance. Pour cela, une sélection backward-forward, retirant les variables superflues au fur et à mesure et se laissant la possibilité de les remettre si besoin, fut utilisée.

Les résultats n'ont pas mis en avant l'intérêt d'une telle sélection, la différence avec une sélection par le seuil de significativité étant ici négligeable. Les performances du modèle se distinguent des autres sur la partie invalidité mais pas sur la partie incapacité. Les résultats des tests de proportionnalité furent finalement cohérents avec la mesure des performances.

Dans la seconde partie de l'étude, les outils techniques et théoriques pour l'utilisation des modèles de durée en assurance de personnes ainsi qu'un modèle spécialement adapté à l'arrêt de travail furent détaillés.

Une implémentation des modèles à partir de réseaux de neurones fut choisie, pour leur facilité à écrire des programmes d'optimisation grâce à leur structure de graphe orienté. Le parcours du graphe dans un sens et dans l'autre permet la mise à jour des poids du réseaux et le calcul du gradient par rétropropagation, en vue d'être utilisé dans un algorithme de descente. Pour faire le lien avec des modèles usuels, à la manière d'un modèle additif généralisé, un perceptron multi-couches revient à relier linéairement les paramètres du programme d'optimisation à une transformée des variables explicatives. Là où la transformée est fixe pour les modèles additifs généralisés, les réseaux de neurones permettent d'optimiser simultanément la transformée de variables utilisée et la partie linéaire. Encore considérés à tort comme des "boites noires", des outils similaires à ceux des forêts d'arbres d'apprentissage sont disponibles pour interpréter les réseaux de neurones. L'implémentation concrète prit la forme d'un package R où l'apprentissage du modèle eut lieu en C++ pour des gains sur les temps de calcul.

Deux techniques d'adaptation des modèles de durée à la censure et la troncature furent comparées, sur données simulées et réelles. L'une consiste à calibrer le modèle sur les données non censurées et à les pondérer afin de simuler la distribution en absence de censure, l'autre à inclure dans le programme d'optimisation la censure et la troncature en travaillant sur la vraisemblance totale. La méthode par pondération est plus simple d'accès mais l'étude sur données simulées révéla une dégradation des performances lorsque le taux de censure augmente et une faible dégradation pour une corrélation entre les variables explicatives et la censure. Néanmoins cette méthode s'avéra plus efficace sur les données réelles, une piste d'interprétation étant que la complexité introduite dans le programme d'optimisation pour la méthode par vraisemblance, sur des cas réels donc plus complexes, peut réduire significativement la qualité de l'apprentissage du modèle.

Le travail se termina sur une proposition de modèle apte à représenter la séparation entre incapacité et invalidité. Il s'agit d'un modèle à mélange de lois de Weibull, où les arrêts sont associés à un risque de court terme ou un risque de long terme. Ce modèle est constitué de deux parties : en traitant les sorties pour non prolongement en invalidité comme des censures, une première partie modélise la durée des arrêts de travail comme si la séparation entre incapacité et invalidité n'existait pas. L'impact de la séparation en deux états est ensuite modélisé dans une seconde partie, reliée à la première puisque nous supposons que la durée des arrêts de risque de court et long terme ne change pas, seules varient les proportions de risque court et de risque long, un arrêt pour le risque court ayant plus de chance de ne pas être prolongé en invalidité. Certaines données du modèle permettent d'éclairer la séparation : la proportion de risque court/long restant à un instant t, la répartition entre risque court et risque long à un instant t ou la probabilité pour un arrêt de ne pas être prolongé en invalidité.

Cependant, l'intérêt d'un tel modèle sur des données réelles, en terme de performances par comparaison grâce au cadre de la première partie en particulier, reste à être démontré.

Executive summary

Work disability duration modelling

Tristan JUDD

February 14^{th} 2018

Collective contracts covering temporary and permanent disability constitute a major part, more than two billions euros, of BTP-Prevoyance's liabilities. Gains in accuracy, even of small order, are important issues for the optimisation of funds management.

Mathematical provisions are usually realised through a frequency-duration model. We focused on the duration part, treated in France by the use of tables representing the survival function or, as done here, the exit rates. Official tables are available, but the construction of experience tables, to adapt to the portfolio of the corporation, is recommended.

Studied durations often span over years and decades, leaving only a fragment observable. Historically, construction of tables was limited to statistical methods specially adapted to the problem. Nowadays, the majority of machine learning methods can be adapted to cover those specificities.

Our study presented the technical and theoretical tools to implement those methods through the use of neural networks, which are still unpopular in actuarial science. Those methods allow more complex way to take in account the french legislation : instead of the modelization in each state of disability (temporal or permanent) imposed by legislation, we tried to model the complete duration. This model would implicitly make the distinction between the two states, classing work stoppings in two categories, a short time span and a long time span risk.

The first part of the study focused on the realisation, and a slight adaptation, of usual rate methods. It allowed us to fix the theoretical and methodological framework for the comparison of those methods with duration models, once converted in rate tables.

Data built included all the required information : beginning and end of observation times, cause of end of observation (censoring or end of disability state), variables used in the construction of the tables (time spent in disability and age at the beginning) and supplementary explanatory covariates. A fast descriptive analysis showed the specificity of the portfolio, in particular the larger proportion of men and workers than in general population.

After a presentation of the methodological and theoretical framework of statistical learning methods, crude rates were calculated through the use of the Kaplan-Meier estimator. Smoothed rates where then determined with two different methods, a cubic spline smoothing and the Whittaker-Henderson smoothing. Smoothing coefficients, responsible of the aspect of the curve, are usually determined visually and through adaptation tests. We preferred a more accurate method, with the determination of the coefficients based on the performance on a validation sample, performance being measured as the difference between observed and predicted rates. Those smoothing methods are theoretically close, and so are the curves of the rates obtained. We used this fact as an occasion to measure the significativity of the performance of ulterior methods : a method is judged statistically better if its performance distinguishes from those of the two smooths.

Cox model was used, as a classical method to take into account supplementary covariates. This model relies on the proportionality hypothesis, meaning that the effect of covariates shouldn't vary over time. The impact of the validity of this hypothesis was measured with the performance, rather than used for an upstream selection of the model. Variable selection is classically done with a significativity test. We preferred here, with a backward-forward selection, to keep the model with the best performance on a validation sample.

Results didn't reveal an interest in using this selection, the difference in performance with a selection through a significativity test being marginal. The model, in comparison with the precedents, shined to determine long term disability but made little difference for short term disability. Proportional hypothesis tests leaded to the same result.

In the second part of the study, tools for the use of duration models in life insurance and a model specific to work disability in France were detailed.

Models were implemented in the form of neural networks, for its facility to write optimisation programs through its structure in oriented graph, allowing the use of retro-propagation to calculate the gradient in descent algorithm. Similarly to more classic methods as a generalised additive model, a multilayer perceptron consists in linking the optimisation program to a linear combination of a transformation of explanatory variables. The optimisation is limited to the linear part in gam, whereas neural networks can be trained to learn simultaneously the linear part and the ideal transformation. Still often qualified as "black boxes", similar tools to the ones used for forest of decision trees are available to interpret neural networks. The concrete implementation took the form of an R package with a C++ learning infrastructure for computational time.

Two methods to adapt duration models to truncation and censoring were compared on simulated and real data. One method consists in calibrating a model only on uncensored data by ponderating them in a way that simulates the uncensored distribution, the other one includes in the optimisation program censoring and truncation through the total likelihood. The ponderation method is more accessible but performances on simulated data showed decreased performance when the proportion of censoring increased and a slight degradation when explanatory variables were correlated to censoring. However, this method performed best on real data, a possible explanation is that the complexity injected in the optimization program for the total likelihood method impacts significatively the learning phase for already complex problem.

We finally proposed a model adapted to the french legislation, separating disability in two states of incapacity and invalidity. We used a weibull mixture modelization, associating work stopping to a short term risk and a long term risk. The final model is in two parts : the first part by treating the end of a claim for non prolongation in invalidity as a censoring source models claim duration as if the distinction between incapacity and invalidity does not exist. A second part takes into account the impact of the separation between those two states. We supposed that claim duration for long and short term risk is not affected by the transition in invalidity, and thus the only variation between the two parts is the repartition between short and long term risk, a short term risk having higher chances of not being prolonged in invalidity. Some information of the models enlight the distinction in two states : short/long term risk proportion given time t, repartition in short and long term risk given time t or the probability for a claim of not being prolonged in invalidity.

Yet, the quality of the model on real data, through the comparison of performances in the statistical framework defined in the first part, has to be demonstrated.