Université de Montréal

Modélisation de l'espérance de vie des clients en assurance

par

Pierre Luc Cyr

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

avril 2013

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Modélisation de l'espérance de vie des clients en assurance

présenté par

Pierre Luc Cyr

a été évalué par un jury composé des personnes suivantes :

Manuel Morales

(président-rapporteur)

Jean-François Angers

(directeur de recherche)

Catherine Paradis-Therrien

(co-directeur)

Pierre Lafaye de Micheaux

(membre du jury)

Mémoire accepté le:

3 avril 2013

SOMMAIRE

Dans ce mémoire, nous proposons une méthodologie statistique permettant d'obtenir un estimateur de l'espérance de vie des clients en assurance. Les prédictions effectuées tiennent compte des caractéristiques individuelles des clients, notamment du fait qu'ils peuvent détenir différents types de produits d'assurance (automobile, résidentielle ou les deux). Trois approches sont comparées. La première approche est le modèle de Markov simple, qui suppose à la fois l'homogénéité et la stationnarité des probabilités de transition. L'autre modèle – qui a été implémenté par deux approches, soit une approche directe et une approche par simulations – tient compte de l'hétérogénéité des probabilités de transition, ce qui permet d'effectuer des prédictions qui évoluent avec les caractéristiques des individus dans le temps. Les probabilités de transition de ce modèle sont estimées par des régressions logistiques multinomiales.

Mots clés: espérance de vie, modèles par chaines de Markov, régression logistique multinomiale, assurance.

SUMMARY

In this master's thesis, we develop a statistical method to estimate the lifetime expectancy of clients in the insurance domain. The forecasts are personnalized according to the clients' own features, the most notable being the fact that they can have any combination of automobile and residential insurance products. Three approaches are compared. The first approach is the simple Markov model which assume homogeneity and stationnarity of the transition probabilities. The other model suggested – which is implemented both by direct computation and by simulation – allows for heterogeneity of the transition probabilities, thus providing forecasts which evolve in time along with the characteristics of the clients. The transitions probabilities are estimated using multinomial logistic regressions.

Keywords : Lifetime expectancy, Markov chain model, multinomial logistic regression, insurance.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

L'assurance est un domaine très compétitif. Les différentes compagnies doivent constamment faire preuve d'innovation afin d'attirer de nouveaux clients. Dans cette optique, des campagnes marketing sont un outil fréquemment utilisé afin d'établir la marque d'une compagnie et de promouvoir ses produits. On dénombre différents outils et mesures qui permettent de quantifier l'impact des campagnes marketing, tels que le taux de vente (comparaison avec un groupe témoin) et la rétention des clients après un an. Toujours dans le soucis d'améliorer le service offert à ses clients, TD Assurance souhaite obtenir une mesure d'espérance de vie de ses clients. Bien que le concept d'espérance de vie soit bien connu, son application dans le cadre de l'assurance pose certains problèmes auxquels nous tenterons d'apporter une réponse dans ce mémoire.

L'objectif principal de ce mémoire est de créer une méthode permettant de mesurer l'espérance de vie des clients en fonction de leurs différentes caractéristiques. Pour la compagnie, les bénéfices seront d'avoir un outil permettant :

- d'évaluer plus précisément les résultats des campagnes marketing ;
- de déterminer les caractéristiques des clients qui restent plus longtemps (âge, nombre de produits, etc.) afin de mieux les cibler dans les campagnes ultérieures.

De plus, d'autres utilisations pourront êtres faites de cet outil :

- mesurer l'impact de certaines initiatives sur la fidélité des clients ;
- mesurer les bénéfices à longs termes dus aux nouvelles affaires.

De nombreuses applications des modèles markoviens ont été effectuées au domaine de l'assurance, que ce soit en modélisation du risque en assurance-vie (Kwon et Jones, 2008) ou encore en planification des provisions (Hesselager,

1994). De plus, diverses méthodes issues de l'analyse de survie ont été appliquées à ce domaine (Czado et Rudolph, 2002). À notre connaissance, la combinaison des modèles markoviens et des méthodes issues de l'analyse de survie dans le but d'obtenir un estimateur du temps espéré de séjour des clients au sein d'une compagnie d'assurance de biens offrant divers produits d'assurance était encore à ce jour inexistante. Pour y parvenir, nous avons basé notre approche sur des modèles proposés dans l'étude de la mobilité des populations, dont les premiers balbutiements remontent à aussi loin que Prais (1955). Ce dernier a modélisé les probabilités de transition entre des classes sociales en Angleterre dans le but d'obtenir le temps moyen passé dans chaque classe sociale. Le modèle par chaînes de Markov choisi dans ce mémoire est celui de McFarland (1970) et de Spilerman (1972), qui ont proposé une façon d'inclure des composantes hétérogènes dans des modèles markovien par le biais de la régression. Les probabilités de transition de notre modèle ont toutefois été estimées à l'aide de la régression logistique multinomiale, qui a été développée en grande partie par Luce (1959) et McFadden (1973).

Du côté de l'entreprise, ce mémoire témoigne de la volonté de la compagnie d'encourager la recherche des étudiants gradués en statistique. En effet, ce mémoire fait suite à d'autres qui ont été réalisés au sein de l'entreprise et qui portent sur le taux d'annulation des polices (Makhzoum, 2002), le profil des clients (Paradis-Therrien, 2007) et sur la fraude (Poissant, 2009).

La méthode proposée pour obtenir un estimateur de l'espérance de vie des clients se décompose en plusieurs étapes : extraction des données, modélisation, calcul de l'espérance de vie, évaluation. D'une façon assez naturelle, l'ordre des chapitres de ce mémoire suit de près les différentes étapes de la méthode proposée, à l'exception de l'extraction des données qui n'est pas couverte. Dans le premier chapitre, nous commençons par présenter quelques notions d'assurances qui sont nécessaires à la compréhension du problème. Les trois chapitres qui suivent couvrent la modélisation effectuée. En effet, le deuxième chapitre traite des modèles par chaînes de Markov, qui sont les fondements de la méthode que nous proposons. Le troisième chapitre couvre

quant à lui la régression logistique multinomiale. Chapeautant les deux chapitres précédents, le quatrième chapitre présente la notion d'espérance de vie et les estimateurs qui seront utilisés afin de l'estimer. Dans le cinquième chapitre, nous appliquons la méthode à notre jeu de données et présentons les résultats de nos modèles, ce qui inclus une brève analyse descriptive des données à notre disposition.

Chapitre 1

NOTIONS EN ASSURANCE

Dans ce chapitre, nous présentons quelques notions liées au domaine de l'assurance en prenant soin de préciser en quoi ces concepts s'appliquent à nos données.

Définition 1.0.1. *L'ensemble des activités d'un client est défini par son* ***compte****. Un compte client peut regrouper plusieurs polices, autant automobiles que résidentielles.*

En général, les clients n'ont qu'un seul compte. Il peut toutefois arriver que certains clients soient titulaires de plus d'un compte. Étant donné qu'il aurait été complexe de tenir compte de la multiplicité des comptes en raison des systèmes d'information utilisés et que cela ne survient que très rarement, nous avons décidé de mener notre analyse au niveau des *comptes*, et non au niveau des *clients*. Tout au long de ce mémoire, nous utiliserons de façon interchangeable les deux termes pour faire référence au compte.

Définition 1.0.2. *La* ***police d'assurance*** *est le contrat liant le client à la compagnie d'assurance. Il y figure la protection prodiguée par la compagnie d'assurance de même que ses limites.*

À un compte peut être lié plusieurs polices d'assurance. À titre d'exemple, si un client assure à la fois une automobile et une habitation, il aura deux polices d'assurances. Notons qu'une police automobile peut être utilisée pour assurer plusieurs automobiles, et qu'une police habitation peut être utilisée pour assurer plusieurs résidences. Il est également possible d'avoir des comptes pour lesquels il y a plusieurs polices automobiles (resp. habitations) servant à assurer plusieurs automobiles (resp. résidences).

Définition 1.0.3. *La **prime** d'une police d'assurance est le montant annuel que le titulaire de la police doit débourser afin de bénéficier de la protection de la police. Il s'agit donc de son prix.*

Chacune des polices d'un compte sont tarifées indépendamment les unes des autres. La tarification d'une police automobile se fait en fonction des caractéristiques des conducteurs et des véhicules assurés alors que la tarification d'une police résidentielle se fait sur la base des résidences assurées. La prime totale liée à un compte est simplement la somme, à un temps donné, de toutes les primes des polices associées au compte du client. La prime totale d'un compte est l'une des variables utilisée dans la modélisation proposée.

Définition 1.0.4. *TD Assurance est un assureur de **groupes**, c'est-à-dire que la compagnie négocie des tarifs avec des associations de diplômés, des associations profesionnelles et des employeurs que ces organismes proposent à leurs membres.*

Pour une même couverture donnée, il est donc possible que deux clients aux caractéristiques similaires aient des primes différentes en fonction du groupe duquel ils sont membres. Le groupe influence la prime de la ou des polices du compte par le biais d'un multiplicateur de prime. Ainsi, différents groupes ont différents facteurs de multiplication en fonction des négociations effectuées entre TD Assurance et les associations ou employeurs.

Définition 1.0.5. *TD Assurance est également un assureur **direct**, c'est-à-dire que la compagnie propose des couvertures à des personnes qui ne font partie d'aucune association de diplômés, associations professionnelles et ne travaillent pas pour un employeur ayant négocié avec la compagnie pour obtenir des tarifs adaptés.*

Définition 1.0.6. *L'ensemble des clients éligibles à un groupe préférentiel de TD Assurance est nommé **marché de l'affinité**. L'ensemble des clients qui n'y est pas éligible constitue le **marché direct**.*

TD Assurance compte plusieurs assureurs (compagnies d'assurances), certains destinés au marché de l'affinité et d'autres au marché direct. La portée de ce mémoire se limite à la modélisation de l'espérance de vie des clients du

marché de l'affinité de TD Assurance, plus particulièrement à la clientèle ontarienne, qui constitue un porte-feuille de polices couramment utilisé au sein de la compagnie pour le développement et l'application de nouveaux modèles.

Définition 1.0.7. *L'**espérance de vie** d'un compte-client est la durée pendant laquelle ce compte possède au moins une police active auprès de la compagnie.*

Ainsi, il n'est pas question ici de temps de vie des individus en termes d'années vécues, mais bien de durée de séjour moyen au sein de la compagnie. Les chapitres qui suivent présentent la méthode que nous proposons afin d'estimer cette espérance de vie. Dans le chapitre suivant, nous introduisons les chaînes de Markov, qui constituent les fondements de notre modélisation.

Chapitre 2

CHAÎNES DE MARKOV À TEMPS DISCRET

Le but du projet étant d'obtenir un estimateur de l'espérance de vie des clients, deux solutions nous sont apparues :

- modéliser le temps de vie directement en choisissant des méthodes utilisées en analyse de survie (modèle de Cox, etc.) ;
- discrétiser le temps, utiliser un modèle multi-états et modéliser les probabilités de transition.

La première solution, quoique intuitivement plus naturelle (nul besoin de discrétiser la variable à expliquer qu'est le temps), posait un nombre certain de problématiques, essentiellement d'ordre pratique. Ces problématiques étaient que :

- les méthodes plus avancées modélisant le temps de vie de façon qui était satisfaisante pour nos besoins, principalement les modèles multi-états tels que le modèle multi-états de Cox (Meira-Machado *et al.*, 2009) sont computationnellement plus lourdes ;
- ces mêmes méthodes possèdent encore peu ou pas d'implémentations dans les logiciels statistiques couramment utilisés ;
- ces méthodes, plus récentes et complexes, sont plus difficiles à comprendre, et le milieu industriel préfère des méthodes fiables et qui ont fait leurs preuves.

Pour toutes ces raisons, nous avons retenu une classe de modèles impliquant une discrétisation de la variable à expliquer, soit les modèles par chaînes de Markov à temps discret. Ces modèles sont relativement simples et fournissent

une base flexible à partir de laquelle des variantes des modèles peuvent être élaborées. C'est donc cette classe de modèles qui est présentée dans ce chapitre. Dans un premier temps, nous présentons des notions de base sur les chaînes de Markov nous permettant de comprendre celles que nous utiliserons. Ensuite, nous présentons différents modèles par chaînes de Markov qui sont d'intérêt pour notre problème : le modèle de Markov simple, le modèle *Mover-Stayer*, le modèle de mobilité de Cornell et le modèle de McFarland-Spilerman.

2.1. INTRODUCTION AUX CHAÎNES DE MARKOV

Avant de présenter la modélisation de nos données par le modèle par chaînes de Markov, nous présentons ici un survol de la théorie des chaînes de Markov à temps discrets. Nous ne présentons que les concepts nécessaires à la compréhension des modèles et prenons soin d'illustrer l'application de ces concepts aux cas qui nous intéressent. Pour plus de détails sur les chaînes de Markov, le lecteur peut consulter Norris (1997).

2.1.1. Définitions

Définition 2.1.1. *Une **chaîne de Markov à temps discrets** est une suite de variables $\{X_n\}_{n\geq 0}$ à valeurs dans un espace d'états fini ou dénombrable J tel que*

$$P\left(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0\right) = P\left(X_{n+1} = j | X_n = i\right) \qquad (2.1.1)$$

où $i, j \in J$ et $n \geq 0$ représente le temps.

La propriété (2.1.1) est nommée propriété markovienne. Lorsque nous parlons des probabilités de transition entre les états, nous utilisons la notation suivante.

Notation 2.1.1. *La probabilité d'aller de l'état i à l'état j, $P(X_n = j | X_{n-1} = i)$ est notée p_{ij}.*

Notation 2.1.2. *La probabilité d'aller de l'état i à l'état j en n pas, $P(X_n = j | X_0 = i)$ est notée $p_{ij}^{(n)}$.*

Exemple 2.1.1. *Considérons une des chaînes de Markov les plus simples, soit le modèle vie-mort. Il y a deux états possibles : l'état de vie et l'état de mort. Une fois la*

mort atteinte, il est impossible d'en sortir. Supposons que la probabilité de décéder à un temps t *donné soit de* $0{,}1$*. Alors la probabilité de demeurer en vie est de* $0{,}9$*. Cette situation a été représentée schématiquement à la figure 2.1, où « 1 » représente l'état*

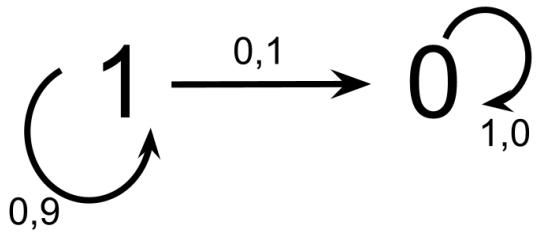


FIG. *2.1: Chaîne de Markov d'un modèle vie-mort*

de vie et « 0 » représente l'état de mort et les valeurs sur les flèches représentent les probabilités d'aller d'un état à un autre.

Pourquoi s'agit-il d'une chaîne de Markov ? En fait, nous avons posé le modèle de telle sorte qu'il satisfasse la condition (2.1.1)*. En effet, on a*

$$P(X_{n+1} = 1 | X_n = 1, X_{n-1} = 1, ..., X_0 = 1) = P(X_{n+1} = 1 | X_n = 1) = 0{,}9$$

et

$$P(X_{n+1} = 0 | X_n = 1, X_{n-1} = 1, ..., X_0 = 1) = P(X_{n+1} = 0 | X_n = 1) = 0{,}1,$$

c'est-à-dire que la probabilité de demeurer en vie ou de décéder ne dépend que du fait que l'individu soit en vie. De même,

$$P(X_{n+1} = 0 | X_n = 0, X_{n-1} = j_{n-1}, ..., X_0 = j_0) = P(X_{n+1} = 0 | X_n = 0) = 1,$$

c'est-à-dire que la probabilité de rester à l'état de mort lorsqu'on y est est certaine et ne dépend pas du passé de l'individu.

Cet exemple aurait pu être utilisé comme base pour un modèle simplifié d'espérance de vie où l'état « vie » serait « avoir au moins une police d'assurance » et l'état mort serait « ne plus avoir de polices d'assurance ». Il ne resterait plus qu'à obtenir les probabilités de transition, par exemple en les estimant à l'aide de régressions logistiques. La chaîne de Markov correspondante est représentée à la figure 2.2.

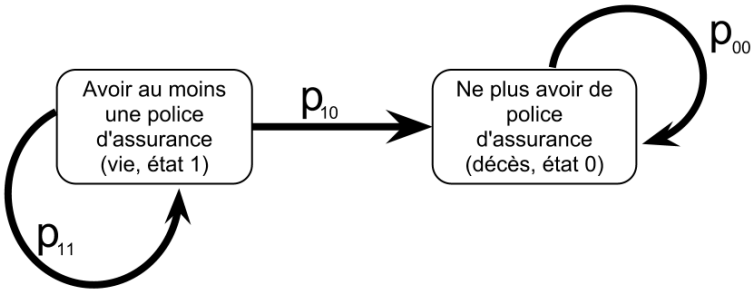


FIG. 2.2: Chaîne de Markov à deux états pour le temps de vie

Cependant, nous avons choisi pour notre méthode une chaîne de Markov plus complexe dont les états sont définis par :

- avoir au moins une police d'assurance automobile et au moins une police d'assurance résidentielle avec la compagnie, que nous définissons comme étant l'état 1 ;
- ne détenir qu'une ou des polices d'assurance automobile avec la compagnie, que nous définissons comme étant l'état 2 ;
- ne détenir qu'une ou des polices d'assurance résidentielle avec la compagnie, que nous définissons comme étant l'état 3 ;
- ne plus détenir aucune police d'assurance avec la compagnie, que nous définissons comme l'état 4. Il s'agit de l'état de « mort », et nous y référerons souvent comme étant la mort (ou le décès) des clients dans ce mémoire.

Quelles sont les transitions possibles entre ces différents états ? Il est possible, pour un client détenant à la fois une police d'assurance automobile et résidentielle de se départir de l'une ou l'autre de ses polices et ce, pour différentes raisons (prime pour cette police trop élevée, se départir du produit en question, etc.). Ainsi, les transitions de l'état 1 à l'état 2 et de l'état 1 à l'état 3 sont possibles (c'est-à-dire que les probabilités de transition sont non nulles). Inversement, il est possible qu'un client ne possédant qu'un seul de ces produits désire s'assurer auprès de la même compagnie (prime intéressante en raison de rabais multi-produits, achat d'un produit que le client n'avait pas, volonté de faire affaire avec une seule compagnie, etc.), de telle sorte que les

transitions de l'état 2 à l'état 1 et de l'état 3 à l'état 1 sont possibles. À partir de chacun des états, il est possible de se rendre à l'état 4 (ne plus avoir aucune police) en un an. Nous n'autorisons pas l'achat de nouvelles polices une fois l'état 4 atteint, bien que cela soit possible autant en théorie qu'en pratique. Deux raisons nous ont mené à cette décision. Premièrement, l'information du client est susceptible de changer lorsqu'il n'est plus au sein de notre compagnie sans qu'elle ne soit mise à jour dans nos systèmes, de telle sorte qu'il aurait été hasardeux de supposer que la dernière information disponible (la même que nous avons utilisé pour modéliser son décès) expliquait son retour dans la compagnie. Deuxièmement, les systèmes d'information et les procédures de création de nouveaux comptes au sein de la compagnie ne nous garantissaient pas qu'à son retour, un client ait le même compte. Pour s'assurer que les clients qui quittent la compagnie ne réintègrent pas notre échantillon, nous avons décidé de ne pas accepter l'entrée de nouveaux clients dans l'étude suivant son début. Finalement, nous n'avons pas non plus autorisé le passage de l'état 2 (n'avoir que des polices automobiles) à l'état 3 (n'avoir que des polices résidentielles) ni le passage inverse. Encore une fois, cela est possible tant en théorie qu'en pratique. Toutefois, cela arrivait en pratique tellement peu souvent qu'il aurait été risqué d'ajuster des modèles : des problèmes de convergence seraient survenus, ou encore, les estimateurs auraient été trop dépendants de l'échantillon choisi (ou, à plus proprement parler, du sous-échantillon des individus effectuant ces transitions). La chaîne de Markov qui a été décrite ci-dessus est représentée schématiquement à la figure 2.3, p.14.

2.1.2. Matrice de transition et matrice de transition à k pas

Une chaîne de Markov peut être représentée par une matrice représentant les probabilités de passer d'un état à un autre. Lorsque le nombre d'états est fini, la matrice sera elle aussi de taille finie.

Définition 2.1.2. *La **matrice de transition** d'une chaîne de Markov à temps discrets est définie par*

$$P = [p_{ij}]_{(i,j)\in J^2} = [P(X_{n+1} = j | X_n = i)]_{(i,j)\in J^2} ,$$

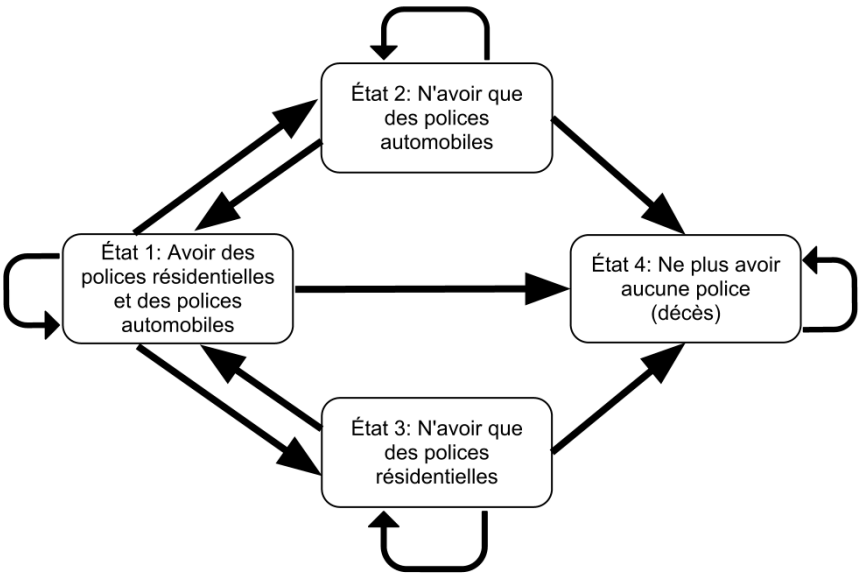


FIG. 2.3: Représentation graphique du modèle multi-états

c'est-à-dire que chaque entrée de la matrice contient la probabilité de passer de l'état i *à l'état* j*. Pour qu'une matrice soit une matrice de transition, il faut également que la somme des éléments d'une ligne soit égale à 1. En effet, chaque élément* j *de la ligne* i *est la probabilité d'aller à* j *à partir de ce* i *donné. Si l'entité ne va pas à un état différent de l'état auquel elle se situe, alors elle restera à ce même état ; ces événements étant mutuellement exclusifs, la somme de leur probabilité doit égaler 1.*

Exemple 2.1.2. *La matrice de transition de la chaîne de Markov du modèle vie-mort décrit auparavant (voir figure 2.2) est*

$$P = \begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix}. \tag{2.1.2}$$

Puisqu'il est impossible de sortir de l'état de mort (état 0), la probabilité p_{00} *est égale à 1 et la probabilité* p_{01} *est égale à 0 de telle sorte qu'en fait*

$$P = \begin{pmatrix} p_{11} & p_{10} \\ 0 & 1 \end{pmatrix}. \tag{2.1.3}$$

Exemple 2.1.3. *La matrice de transition de la chaîne de Markov représentée à la figure 2.3 est*

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & 0 & p_{24} \\ p_{31} & 0 & p_{33} & p_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.1.4}$$

Les matrices de transition sont très utiles pour effectuer des calculs impliquant des chaînes de Markov. En ce qui nous concerne, nous les utiliserons essentiellement afin de déterminer le temps requis pour se rendre à l'état 4 (décès) à partir de chacun des autres états. Pour ce faire, nous devons définir la puissance d'une matrice.

Définition 2.1.3. *Soit une matrice* $P = ||p_{ij}||$ *de taille* $n \times n$. *Le carré de la matrice* P, *noté* P^2, *est obtenu en effectuant le produit matriciel de la matrice avec elle-même :*

$$P^2 = ||p_{ik}^{(2)}|| = ||\sum_{j=1}^{n} p_{ij}p_{jk}|| = PP. \tag{2.1.5}$$

Donc, la deuxième puissance d'une matrice n'est que le cas particulier du produit matriciel d'une matrice carrée avec elle-même. Dans le même ordre d'idée, nous pouvons définir la k^e puissance d'une matrice carrée.

Définition 2.1.4. *Soit une matrice* $P = ||p_{ij}||$ *de taille* $n \times n$. *La* k^e *puisance de la matrice* P, *noté* P^k, *est obtenue en effectuant le produit matriciel de la matrice avec elle-même* k *fois :*

$$P^k = ||p_{ij}^{(k)}|| = \overbrace{PP \dots P}^{k \text{ fois}}. \tag{2.1.6}$$

Par définition, on pose la puissance 0 *comme étant la matrice identité* I_n.

Il est possible d'utiliser cette définition de puissance d'une matrice afin de répondre à la question : « combien de temps faut-il afin de se rendre à l'état 4 à partir de chacun des autres états ? ». En effet, dans une chaîne de Markov à temps discret, la probabilité de passer de l'état i à l'état j, p_{ij}, est en fait la probabilité d'aller de i à j **dans une période de temps** t **donnée**, nommée la période. Dans le problème qui nous est d'intérêt, la période est d'une année. Fait intéressant et extrêmement utile, lorsque nous multiplions une matrice de

transition avec elle-même – c'est-à-dire lorsque nous l'élevons au carré – nous obtenons une matrice de transition à deux pas. En d'autres mots, la matrice résultante nous donne la probabilité d'aller de l'état i à l'état j en *deux* périodes de temps. De la même façon, la k^e puissance d'une matrice de transition nous donne la probabilité d'aller de l'état i à l'état j en k pas et nous notons $p_{ij}^{(k)}$ cette probabilité.

Exemple 2.1.4. *La matrice de transition donnant les probabilités de transition en deux pas du modèle vie-mort est*

$$P^2 = \begin{pmatrix} p_{11} & p_{10} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} p_{11} & p_{10} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} p_{11}p_{11} & p_{11}p_{10} + p_{10} \\ 0 & 1 \end{pmatrix}. \tag{2.1.7}$$

Exemple 2.1.5. *La matrice de transition donnant les probabilités de transition en deux pas du modèle à quatre états de la figure 2.3 est*

$$\begin{aligned} P &= \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & 0 & p_{24} \\ p_{31} & 0 & p_{33} & p_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & 0 & p_{24} \\ p_{31} & 0 & p_{33} & p_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} p_{11}p_{11}+p_{12}p_{21}+p_{13}p_{31} & p_{11}p_{12}+p_{12}p_{22} & p_{11}p_{13}+p_{13}p_{33} & p_{11}p_{14}+p_{12}p_{24}+p_{13}p_{34}+p_{14} \\ p_{21}p_{11}+p_{22}p_{21} & p_{21}p_{12}+p_{22}p_{22} & p_{21}p_{13} & p_{21}p_{14}+p_{22}p_{24}+p_{24} \\ p_{31}p_{11}+p_{33}p_{31} & p_{31}p_{12} & p_{31}p_{13}+p_{33}p_{33} & p_{31}p_{14}+p_{33}p_{34}+p_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \end{aligned}$$

Il est intéressant de constater que chaque élément de la matrice de transition à deux pas peut se lire comme étant la somme des probabilités de toutes les façons possibles de se rendre de l'état i à l'état j. Ainsi, le premier élément, $p_{11}p_{11}+p_{12}p_{21}+p_{13}p_{31}$ signifie que la probabilité de se rendre de l'état 1 à l'état 1 en deux pas est la somme des probabilités des événements suivant :

– rester à l'état 1 au premier pas et y rester également au second pas ;

– aller à l'état 2 au premier pas et revenir à l'état 1 au second pas ;

– aller à l'état 3 au premier pas et revenir à l'état 1 au second pas.

Pour des puissances élevées, la multiplication matricielle devient rapidement fastidieuse, surtout lorsque le nombre d'états est grand. Lorsque qu'il est

possible de diagonaliser la matrice de transition – c'est-à-dire lorsqu'elle admet autant de valeurs propres distinctes que d'états – le calcul de P^k devient beaucoup plus simple en utilisant le fait que $P^k = Q^{-1}A^kQ$, où A est diagonale. Norris (1997) propose la méthode suivante afin de calculer les $p_{ij}^{(k)}$ pour n'importe quelle chaîne de Markov à M états (donc lorsque la matrice de transition est de taille $M \times M$) :

- Calculer les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_M$ de la matrice de transition P en trouvant les solutions de l'équation

$$0 = det(\lambda I_M - P). \tag{2.1.8}$$

- Lorsque les valeurs propres sont toutes distinctes, résoudre, pour chacune des combinaisons de i et de j, le système d'équations

$$p_{ij}^{(k)} = \alpha_1\lambda_1^k + \alpha_2\lambda_2^k + \cdots + \alpha_M\lambda_M^k, \tag{2.1.9}$$

  où les constantes $\alpha_1, \alpha_2, \ldots, \alpha_M$ dépendent de i et j. Il faut obtenir autant d'équations qu'il y a de valeurs propres en trouvant $p_{ij}^{(0)}, p_{ij}^{(1)}, \ldots$ Pour ce faire, il peut être nécessaire de trouver les premières puissances de P en effectuant les multiplications matricielles.
- Lorsqu'une valeur propre est répétée, il faut modifier la forme générale de $p_{ij}^{(k)}$ pour en tenir compte. Par exemple, si λ_1 est répétée une fois, alors la forme générale de $p_{ij}^{(k)}$ devient

$$p_{ij}^{(k)} = (k\alpha_1 + b)\lambda_1^k + \alpha_2\lambda_2^k + \cdots + \alpha_M\lambda_M^k. \tag{2.1.10}$$

Appliquons maintenant cette méthode à nos deux chaînes de Markov.

Exemple 2.1.6. *Considérons la chaîne de Markov dont la matrice de transition est*

$$P = \begin{pmatrix} 0{,}9 & 0{,}1 \\ 0 & 1 \end{pmatrix}. \tag{2.1.11}$$

Les valeurs propres de P *sont obtenues en trouvant la solution de l'équation caractéristique*

$$0 = det(\lambda I_M - P) = (\lambda - 0{,}9)(\lambda - 1) \tag{2.1.12}$$

qui nous donne, dans ce cas, $\lambda_1 = 0{,}9$ *et* $\lambda_2 = 1{,}0$ *étant donné que les valeurs propres d'une matrice triangulaire sont les éléments diagonaux. Ensuite, nous avons que* $p_{11}^{(k)}$ *sera de la forme*

$$p_{11}^{(k)} = \alpha_1 (0{,}9)^k + \alpha_2. \tag{2.1.13}$$

Nous pouvons écrire les deux premières équations, c'est-à-dire celles pour $k = 0$ *et* $k = 1$*, afin d'obtenir un système à deux équations et deux inconnues que nous pouvons résoudre :*

$$1 = p_{11}^{(0)} = \alpha_1 + \alpha_2$$
$$0{,}9 = p_{11}^{(1)} = 0{,}9\alpha_1 + \alpha_2$$

et qui a pour solution $\alpha_1 = 1$ *et* $\alpha_2 = 0$*. Ainsi, la forme générale de* $p_{11}^{(k)}$ *est donnée par*

$$p_{11}^{(k)} = 0{,}9^k. \tag{2.1.14}$$

Il est possible d'effectuer le même calcul pour les autres éléments de la matrice de transition ou, dans ce cas-ci, d'utiliser la propriété voulant que la somme des lignes d'une matrice de transition soit égale à 1 pour obtenir

$$p_{10}^{(k)} = 1 - 0{,}9^k \tag{2.1.15}$$

de telle sorte que la matrice de transition à k *pas est*

$$P^k = \begin{pmatrix} 0{,}9^k & 1-0{,}9^k \\ 0 & 1 \end{pmatrix}. \tag{2.1.16}$$

Exemple 2.1.7. *Supposons que, dans la chaîne de Markov à quatre états donnée précédemment, la matrice de transition soit*

$$P = \begin{pmatrix} 0{,}8 & 0{,}1 & 0{,}05 & 0{,}05 \\ 0{,}2 & 0{,}7 & 0 & 0{,}1 \\ 0{,}2 & 0 & 0{,}6 & 0{,}2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Nous pouvons résoudre l'équation caractéristique

$$0 = \det(\lambda I_M - P) = \begin{vmatrix} \lambda - 0{,}8 & 0{,}1 & 0{,}05 & 0{,}05 \\ 0{,}2 & \lambda - 0{,}7 & 0 & 0{,}1 \\ 0{,}2 & 0 & \lambda - 0{,}6 & 0{,}2 \\ 0 & 0 & 0 & \lambda - 1 \end{vmatrix}$$

afin de trouver les valeurs propres $\lambda_1 = 0{,}532487$, $\lambda_2 = 0{,}646081$, $\lambda_3 = 0{,}921432$ *et* $\lambda_4 = 1$. *À titre d'exemple, nous calculons le terme* $p_{11}^{(k)}$. *Nous savons qu'il est de la forme*

$$\begin{aligned} p_{11}^{(k)} &= \alpha_1\lambda_1^k + \alpha_2\lambda_2^k + \alpha_3\lambda_3^k + \alpha_4\lambda_4^k \\ &= \alpha_1(0{,}532487)^k + \alpha_2(0{,}646081)^k + \alpha_3(0{,}921432)^k + \alpha_4. \end{aligned}$$

Il nous faut obtenir les valeurs de $p_{11}^{(0)}$, $p_{11}^{(1)}$, $p_{11}^{(2)}$ *et* $p_{11}^{(3)}$. *Bien que les deux premières quantités soient triviales, il nous faut effectuer les multiplications matricielles* $P^2 = PP$ *et* $P^3 = PP^2$ *afin d'obtenir les deux dernières. Lorsque nous effectuons ces calculs, nous obtenons*

$$P^2 = \begin{pmatrix} 0{,}67 & 0{,}15 & 0{,}07 & 0{,}11 \\ 0{,}30 & 0{,}51 & 0{,}01 & 0{,}18 \\ 0{,}28 & 0{,}02 & 0{,}37 & 0{,}33 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

et

$$P^3 = \begin{pmatrix} 0{,}5800 & 0{,}1720 & 0{,}0755 & 0{,}1725 \\ 0{,}3440 & 0{,}3870 & 0{,}0210 & 0{,}2480 \\ 0{,}3020 & 0{,}0420 & 0{,}2360 & 0{,}4200 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

de telle sorte que $p_{11}^{(2)} = 0{,}67$ *et* $p_{11}^{(3)} = 0{,}58$*. Nous avons donc le système de quatre équations à quatre inconnues*

$$\begin{aligned}
1 = p_{11}^{(0)} &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \\
0{,}8 = p_{11}^{(1)} &= \alpha_1(0{,}532487) + \alpha_2(0{,}646081) + \alpha_3(0{,}921432) + \alpha_4 \\
0{,}67 = p_{11}^{(2)} &= \alpha_1(0{,}532487)^2 + \alpha_2(0{,}646081)^2 + \alpha_3(0{,}921432)^2 + \alpha_4 \\
0{,}58 = p_{11}^{(3)} &= \alpha_1(0{,}532487)^3 + \alpha_2(0{,}646081)^3 + \alpha_3(0{,}921432)^3 + \alpha_4
\end{aligned}$$

qui, une fois résolu, nous donne $\alpha_1 = 0{,}255972$*,* $\alpha_2 = 0{,}0794364$*,* $\alpha_3 = 0{,}664592$ *et* $\alpha_4 = 0$*. Ainsi,*

$$p_{11}^{(k)} = 0{,}255972\lambda_1^k + 0{,}0794364\lambda_2^k + 0{,}664592\lambda_3^k.$$

Il est possible d'obtenir les autres termes de la matrice de transition à k *pas en procédant de la même façon.*

Dans cette section, nous avons vu comment obtenir la matrice de transition à k pas. Cette quantité nous sera utile au chapitre 4, chapitre dans lequel nous expliquons comment l'estimateur de l'espérance de vie est obtenu dans les différents modèles utilisés.

2.2. Modélisation statistique par chaîne de Markov : modèle de Markov simple

Maintenant que nous avons donné les concepts de base des chaînes de Markov, nous pouvons nous attarder à la modélisation statistique de ces chaînes. En d'autres termes, nous voyons maintenant des modèles statistiques, comportant leurs lots d'hypothèses, qui permettent d'estimer les probabilités de transition des modèles dans un cadre markovien. Le premier tel modèle que nous voyons est le modèle de Markov simple. Ce modèle, qui est l'un des trois utilisés dans ce mémoire, consiste simplement à estimer les probabilités de transition par les fréquences empiriques.

2.2.1. Hypothèses

Lorsque nous utilisons les fréquences empiriques comme estimateurs des probabilités de transitions, nous supposons que les mécanismes dictant les comportements des individus de notre population respectent certaines conditions ; ce sont à ces conditions que nous nous intéressons ici. Ces hypothèses sont au nombre de trois : propriété de Markov, stationnarité et homogénéité.

La première de ces hypothèses est sans surprise : tout modèle de Markov doit satisfaire à la propriété markovienne, qui stipule que les probabilités d'un individu donné au temps k ne dépend que de l'emplacement de cet individu dans le système à ce temps k donné et non de ses états précédents. Pour illustrer la situation, considérons un exemple issu de l'étude de la mobilité des populations, soit celui des mouvements entre différentes régions géographiques d'une génération à une autre. Comme le souligne Hodge (1966), l'utilisation du modèle de Markov simple afin de modéliser les mouvements de populations d'une génération à une autre impose la contrainte que les probabilités de transition ne dépendent que de l'influence des parents sur les enfants. Ainsi, l'influence des *grands-parents* sur les transitions de leurs petits-enfants est supposée nulle. Dans le cas qui nous intéresse, les clients demeurent les mêmes d'un état à l'autre. Il est toutefois permis de se demander s'il est réaliste de supposer que la transition à un temps donné ne dépende que de l'état actuel. Nous pensons que oui, et voici pourquoi. Un client peut posséder plusieurs polices. Lorsque vient le temps de renouveler une ou plusieurs de ses polices (changement possible d'état), nous pensons que le client se demande *au moment où il fait ce choix* si le prix proposé lui convient en fonction des produits qu'il assure et des couvertures proposées. Il serait audacieux de prétendre qu'une entité économique telle qu'un client, qui prend des décisions économiquement rationnelles, tienne compte de ses primes et couvertures passées : si elle veut minimiser sa prime, elle doit le faire au moment de prendre sa décision. C'est pourquoi nous pensons que l'hypothèse de Markov, qui est faite dans chacun des modèles qui ont été ajustés, est réaliste. Mais qu'arriverait-il si cette hypothèse n'était pas satisfaite ? Toujours selon Hodge (1966), qui a mené une étude

empirique, il est possible d'observer l'impact du non-respect de cette hypothèse en comparant les résultats observés versus ceux prédits par le modèle puisque l'hypothèse de Markov signifie mathématiquement que $P^2 = PP$: tout écart entre les fréquences empiriques et les probabilités de transition à deux pas devrait indiquer une déviation de cette hypothèse.

La deuxième hypothèse est celle de stationnarité des probabilités de transition. En estimant les probabilités de décès par les fréquences empiriques, nous imposons cette condition à nos individus – puisque la quantité mathématique par laquelle nous estimons les probabilités la respecte. Mais quelle est, exactement, cette condition ? En fait, par stationnarité, il est impliqué que les probabilités de transition demeurent fixes à chaque temps en raison de l'absence de changements structurels dans le mécanisme qui est modélisé. Inversement, des changements structurels dans les probabilités de transition causeraient une non stationnarité de ces probabilités. En assurance, ces changements structurels pourraient être dus à l'évolution du marché, qui modifie les « règles du jeu » : arrivée (ou départ) de compétiteurs ou encore de modifications dans les comportements des consommateurs liés à de nouvelles législations. Suite à ces propos, il ne semble pas très réaliste d'effectuer l'hypothèse de stationnarité dans le cadre assurantiel. En effet, il serait souhaitable de pouvoir tenir compte de ces différents éléments dans la modélisation de l'espérance de vie des clients. Il importe toutefois de conserver en mémoire que l'objectif étant d'obtenir une mesure d'espérance de vie pour les temps *futurs*, effectuer une modélisation supposant la non stationnarité relève de l'utopie étant donné que la prédiction de l'évolution du marché de l'assurance à long terme (il est ici question de plusieurs dizaines d'années) relève davantage des arts divinatoires que de la statistique. Cette hypothèse – ou contrainte – sera donc effectuée dans l'ensemble des modélisations qui seront effectuées dans ce mémoire ; tout estimateur d'espérance de vie obtenu devra donc être interprété comme l'espérance de vie en supposant que les conditions du marché demeurent les

mêmes sur toutes les années de projection. Toutefois, puisque le temps est utilisé comme variable explicative dans le modèle, l'impact de la non stationnarité devrait être moindre.

La troisième hypothèse est celle d'homogénéité des probabilités de transition. Cette hypothèse veut que les probabilités de transition soient les mêmes pour tous les individus. Cette hypothèse est plausible lorsque nous savons que les comportements des entités modélisées sont similaires et ne sont pas dictés par des différences propres à ces entités. L'assurance automobile étant obligatoire par la loi et l'assurance résidentielle étant fortement recommandée, les clients de notre étude forment un tout hétérogène ; ces différentes caractéristiques pouvant influencer leurs comportements en matière d'assurance, il n'est pas rationnel d'effectuer cette hypothèse.

Nous constatons que les hypothèses du modèle de Markov simple ne correspondent pas à ce que le contexte de la modélisation de l'espérance de vie des clients en assurance semble exiger. C'est pourquoi nous avons cherché des modèles plus élaborés présentant des caractéristiques plus près de ce qui est souhaitable pour notre problème. Ces modèles sont présentés aux sections 2.3, 2.4 et 2.5. Nous avons néanmoins modélisé l'espérance de vie en utilisant le modèle de Markov simple, et ce, pour deux raisons. D'une part, il est très simple à ajuster. D'autre part, cela nous donne une modélisation baromètre permettant de comparer les performances des autres modèles qui ont été ajustés.

2.2.2. Estimation des probabilités de transitions

Tel que mentionné précédemment, le modèle de Markov simple estime les probabilités de transition par les fréquences empiriques : en effet, le nombre d'individus qui sont passés de l'état i à l'état j divisé par le nombre de personnes partant de l'état i donne un estimateur de p_{ij}. En fait, Anderson et Goodman (1957) ont démontré que ces quantités sont les estimateurs du maximum de vraisemblance. Ils ont démontré ce résultat à la fois pour le cas des probabilités de transitions stationnaires et non stationnaires, c'est-à-dire qui

évoluent dans le temps. Notons que lorsque nous supposons la non stationnarité d'un modèle de Markov simple, il faut pouvoir observer les changements d'états sur l'ensemble des temps pour lesquels des estimateurs des probabilités de transition sont souhaités.

Comme précédemment, nous supposons une chaîne de Markov à m états et nous avons T années d'observation disponibles pour la variable temps $t = 0, 1, \ldots, T$. Nous utilisons la notation $p_{ij}(t)$ pour représenter la probabilité d'aller à l'état j au temps t sachant un départ de i au temps $t-1$. Le cas où $p_{ij}(0) = p_{ij}(1) = \ldots p_{ij}(T) = p_{ij}$ est simplement celui où les probabilités de transitions sont stationnaires. Nous notons $n_{ij}(t)$ le nombre d'individus qui sont à l'état j au temps t et qui étaient à l'état i au temps $t-1$. On pose $n_{ij} = \sum_{t=1}^{T} n_{ij}(t)$, qui représente le nombre total d'individus qui sont passés de i à j pendant la durée de l'étude. Alors, dans le cas où les probabilités de transitions sont stationnaires, l'estimateur du maximum de vraisemblance de p_{ij} est

$$\hat{p}_{ij} = \frac{\sum_{t=1}^{T} n_{ij}(t)}{\sum_{k=1}^{m} \sum_{t=1}^{T} n_{ik}(t)} \quad i, j \in J \tag{2.2.1}$$

alors que lorsque les probabilités de transitions sont non stationnaires, l'estimateur du maximum de vraisemblance de p_{ij} est plutôt

$$\hat{p}_{ij}(t) = \frac{n_{ij}(t)}{\sum_{k=1}^{m} n_{ik}(t)} \quad i, j \in J. \tag{2.2.2}$$

Ces expressions sont naturelles et ne constituent qu'une formalisation mathématique de ce que nous avons mentionné au début de cette section. Nous attirons toutefois l'attention sur le fait que, dans le cas où les probabilités de transitions sont stationnaires, nous regardons le nombre total d'individus qui sont passés de l'état i à l'état j pendant l'étude par rapport au nombre total d'individus qui se sont retrouvés à l'état i pendant l'étude (et lorsqu'un individu y est resté plusieurs temps, il est compté autant de fois que le nombre de temps qu'il y est demeuré). Dans le problème qui nous intéresse, nous ne disposons que d'une année d'observation, c'est-à-dire que $T = 1$.

2.3. Le modèle Mover-Stayer

Différents auteurs ont proposé différentes variations du modèle de Markov simple, dont les hypothèses peuvent être contraignantes ou inappropriées. L'une de ces alternatives est le modèle *Mover-Stayer*, qui suppose que la population est divisée en deux types d'individus : ceux qui changent d'état (*movers*) et ceux qui demeurent au même état (*stayers*). Ce faisant, nous supposons une certaine forme d'hétérogénéité de la population en lieu et place de l'homogénéité du modèle de Markov simple. Les autres hypothèses (propriété markovienne, stationnarité) demeurent quant à elles inchangées. Les paramètres qui doivent être estimés dans le modèle *Mover-Stayer* sont la proportion d'individus qui changent d'état de même que les probabilités de transition de ces individus. Goodman (1961) montre des résultats théoriques sur divers estimateurs utilisés dans ce modèle.

Ce modèle aurait pu être intéressant dans le cadre de la modélisation de l'espérance de vie des clients. En effet, il ne serait pas aberrant de supposer qu'une certaine partie de la population est fidèle à la compagnie coûte que coûte et que l'autre partie finira par quitter la compagnie. Toutefois, l'adoption de ce modèle reviendrait à accepter *a priori* que l'espérance de vie soit infinie, à moins de forcer artificiellement le décès des clients au-delà d'un certain seuil de temps de vie. Pour cette raison, nous avons cherché un modèle plus élaboré que le modèle *Mover-Stayer* et ne l'avons donc pas modélisé.

2.4. Le modèle de mobilité de Cornell

Une autre variation du modèle de Markov simple est le modèle de mobilité de Cornell. Ce modèle se fonde sur la prémisse que plus longtemps un individu séjourne à un état donné, plus grande est sa probabilité de demeurer à cet état pour une autre unité de temps (McGinnis, 1967). McFarland (1970) souligne que cette variation semble violer les trois hypothèses du modèle de Markov, en ce sens que les les probabilités de transitions changent dans le temps (apparence de non stationnarité), que la probabilité de transition d'une entité dépend de son passé (apparence de non respect de la propriété de Markov) et

que des entités qui sont à un même état peuvent avoir des probabilités de transition différentes (apparence d'hétérogénéité) mais qu'une simple reformulation du modèle prouve le contraire : il suffit de créer autant d'états qu'il y a de temps de séjours à ces états pour que les trois hypothèses soient à nouveau satisfaites. Nous pouvons représenter les nouveaux états comme le couple (n, t) où $n \in \eta$ représente les états possibles de la chaîne de Markov et $t = 1, 2, \ldots$ représente le temps consécutif passé à cet état. Un problème pratique surgit alors : la quantité d'états possibles est très grande, voire infinie, ce qui peut poser des problèmes d'estimation. Pour cette raison, et parce que nous désirons avoir un modèle permettant de tenir compte des caractéristiques individuelles de nos clients, nous avons préféré opter pour le modèle plus élaboré qui est présenté à la section suivante.

2.5. MODÈLE DE MCFARLAND-SPILERMAN

Dans l'optique d'obtenir une plus grande flexibilité, des modèles permettant de tenir compte de l'hétérogénéité inhérente à des ensembles de données ont été développés. Dans cette optique, McFarland (1970) discute d'un modèle markovien dans lequel chaque *individu* est une chaîne de Markov ; bien qu'il mette en garde qu'une fois agrégé au niveau de la population, le processus ne soit pas nécessairement markovien, il précise qu'il admet une distribution stationnaire dont la matrice de transition est de la forme

$$Q = N_0^{-1} \sum_c N_c P_c^{(1)}(0), \tag{2.5.1}$$

où $P_c^{(1)}(0)$ est la matrice de transition à un pas au temps 0 de l'individu c, N_c est la matrice carrée comportant un 1 à l'élement diagonal correspondant à l'état de départ de l'individu c et des zéro ailleurs et où N_0^{-1} est la matrice diagonale obtenue en inversant la matrice $N = \sum_c N_c$. McFarland, bien que ne spécifiant pas suffisamment son modèle pour qu'il puisse être estimé, indique en quoi il pourrait être utile. En effet, il explique que de l'hétérogénéité peut résulter une évolution des probabilités de transition dans le temps sans pour autant qu'il n'y ait de changements structurels. Par exemple, si les clients plus âgés sont

plus fidèles et que la population modélisée vieillit, alors nous pouvons nous attendre à ce que les probabilités de décès diminuent dans le temps : l'effet de la caractéristique « âge » ne change pas dans le temps ; c'est la prévalence des valeurs plus élevées dans la population modélisée qui augmente. Il s'agit d'une façon originale de tenir compte d'une partie de l'évolution des probabilités de transition dans le temps. Dans notre cas, il s'agit exactement de la partie dont il est possible de tenir compte, les changements structurels du marché ne pouvant être anticipés sur la base des données disponibles. Cette voie, qui revient à effectuer l'hypothèse de stationnarité mais non celle de l'homogénéité, est la voie de prédilection qui a été choisie pour la modélisation de nos données ; c'est dans cette méthode que nous fondons le plus d'espoirs étant donné que c'est celle dont les hypothèses semblent les plus près et les plus réalistes pour notre notre situation. Plus particulièrement, la méthode utilisée sera analogue à celle proposée par Spilerman (1972) qui, pour introduire le concept d'hétérogénéité dans le cadre markovien, se base sur le principe que les différences entre les matrices de transition des entités sont dues à des différences entre les caractéristiques individuelles de ces entités. La méthode la plus usitée pour traiter ce genre de situations étant la régression, c'est celle qu'il a choisi. Dans un premier temps, Spilerman obtient l'expression théorique de l'estimateur de P en fonction des caractéristiques individuelles. Les $\hat{p}_{ij}$ sont estimés par des régressions linéaires simples dans lesquelles la variable explicative est codée 0 ou 1. Cette matrice, que nous noterons $\hat{P}_X^{(1)}$ afin de se rappeler qu'elle doit être évaluée aux valeurs de la matrice des variables explicatives X, peut être utilisée afin d'obtenir un portrait de l'influence de chacune des caractéristiques (dont les coefficients sont représentés dans les $\hat{p}_{ij}$) sur les mouvements entre les états. Par ailleurs, puisque l'expression obtenue est conditionnelle à la matrice des variables explicatives X, il est possible d'obtenir une expression numérique de $\hat{P}^{(1)}$ pour des valeurs *hypothétiques* des variables explicatives. L'utilité la plus immédiate est l'obtention de matrices de transition personnalisées pour différents sous-ensembles de la population étudiée. Dans un second

temps, Spilerman projette son estimateur de $\hat{P}^{(1)}$ pour les temps futurs en tenant compte de l'évolution des caractéristiques individuelles dans le temps, obtenant ainsi un estimateur de $\hat{P}^{(k)}$. Notons $\hat{P}^{(q)}(t)$ la matrice de transition à q pas au temps t. Nous avions vu que dans le cas homogène, un estimateur de $P^{(k)}$ est donné par

$$\hat{P}^{(k)} = \prod_{t=0}^{k} \hat{P}^{(1)}(t) = \hat{P}^{k},$$

où la dernière égalité est valide lorsque les probabilités de transitions sont stationnaires. Dans le cas hétérogène, il faut commencer par obtenir les matrices de transition à un pas individuelles des c individus à chacun des temps t, $M_c^{(1)}(t)$, en évaluant $\hat{P}_X^{(1)}$ avec les caractéristiques de l'individu c, X_c, au temps t. Il faut donc obtenir $c \times t$ telles matrices. Un exemple typique de caractéristique qui évolue dans le temps est l'âge. Ensuite, un estimateur de $\hat{P}^{(k)}$ au niveau de la population peut être obtenu en agrégeant les matrices de transition des c individus à chacun des temps :

$$\hat{P}^{(k)} = N^{-1} \sum_c N_c \left(\prod_{t=0}^{k-1} M_{ct}(1) \right), \tag{2.5.2}$$

où N_c est la matrice ayant un 1 sur la diagonale correspondant à l'état initial de l'individu c et des 0 ailleurs et où N^{-1} est l'inverse de la matrice $N = \sum_c N_c$.

Nous avons mentionné que nous avons appliqué une méthode similaire à celle de Spilerman pour la modélisation du modèle de Markov utilisé pour l'obtention de notre mesure d'espérance de vie. La seule différence est que nous avons modélisé les probabilités de transitions à l'aide de régressions logistiques multinomiales plutôt que par des régressions linéaires. C'est pourquoi le chapitre suivant est consacré à ce type de régression.

2.6. APPROCHE PAR SIMULATIONS

En plus du modèle de Markov simple et du modèle de McFarland-Spilerman, une autre modélisation a été utilisée pour obtenir un estimateur de l'espérance de vie. À l'instar des deux premières, cette approche utilise la chaîne de Markov représentée à la figure 2.3, p.14. Tout comme pour le modèle de

McFarland-Spilerman, des régressions logistiques sont utilisées afin d'obtenir les probabilités de transition en fonction des caractéristiques individuelles. Toutefois, plutôt que d'utiliser la formule (2.5.2) afin de projeter la matrice de transition aux temps futurs, nous utilisons une approche par simulations. En effet, une fois les régressions logistiques multinomiales obtenues, nous calculons, pour chacun des clients de l'étude, ses probabilités de survie à partir de son état initial à l'aide de ces régressions. Nous partitionnons l'espace de probabilités $[0, 1]$ en sous-espaces qui dépendent de ses probabilités de survies ; ces sous-espaces individualisés sont $[0, p_{11}]$, $]p_{11}, p_{11} + p_{12}]$, $]p_{11} + p_{12}, p_{11} + p_{12} + p_{13}]$, $]p_{11} + p_{12} + p_{13}, 1]$ lorsque le client est à l'état 1, $[0, p_{21}]$, $]p_{21}, p_{21} + p_{22}]$, $]p_{21} + p_{22}, 1]$ lorsqu'il est à l'état 2 et $[0, p_{31}]$, $]p_{31}, p_{31} + p_{33}]$, $]p_{31} + p_{33}, 1]$ lorsqu'il est à l'état 3. Ensuite, nous générons des variables aléatoires uniformes entre 0 et 1. Ce sont ces variables aléatoires qui nous permettent de classer aléatoirement le client à un état futur. Cette façon de procéder nous permet d'obtenir un « chemin aléatoire » à travers les états de la chaîne de Markov pour chacun des clients de notre étude. Nous pouvons donc affirmer que cette approche utilise deux types de « simulations ». La première est celle au niveau individuel : pour chacun des individus, nous simulons les chemins qu'il peut emprunter à l'aide d'une classification aléatoire (mais proportionnelle à ses probabilités de transition). La deuxième, nécessaire afin d'annuler la composante aléatoire des chemins individuels, consiste à répéter la simulation des chemins pour chacun des individus un certain nombre de fois.

Nonobstant les erreurs de simulations, ces deux méthodes devraient donner les mêmes résultats. Il s'agit en effet dans les deux cas de la même chaîne de Markov et des mêmes probabilités de transitions. Les mêmes hypothèses sont effectuées. En effet, dans l'approche par simulations, l'hypothèse de stationnarité découle de la réutilisation des régressions logistiques multinomiales à chacun des temps prédits. Il importe de constater que ces régressions sont modélisées sur la première année : il est donc supposé que l'effet des variables n'évolue pas dans le temps. L'hypothèse d'hétérogénéité découle évidemment

de l'utilisation même de ces régressions logistiques, qui fournissent des probabilités de transitions personnalisées à chacun des clients. Mais alors, pourquoi implémenter ces deux approches, si elles devraient donner le même résultat ? Initialement, nous avions prévu que la méthode de McFarland-Spilerman allait pouvoir être calculée pour l'ensemble des individus de l'étude, ce que les capacités techniques de la distribution du logiciel R disponible sur les postes de travail de la compagnie ne permettait pas d'effectuer au moment où ce mémoire a été rédigé. Nous avons donc dû nous limiter à des sous-ensembles aléatoires des échantillons d'entraînement et de validation de taille 12 000. De plus, il n'était pas évident de déterminer, *a priori*, laquelle des deux méthodes suivantes serait la plus rapide : le calcul matriciel impliquant des quantités importantes de matrices ou la simulation répétée des chemins empruntés par les individus. Dans ce contexte, l'approche par simulations présentait l'avantage de pouvoir choisir la quantité désirée de simulations par individus en fonction de la précision voulue et du temps disponible pour l'obtention de l'estimateur de l'espérance de vie. Finalement, un des objectifs pratiques de ce mémoire était d'obtenir une méthode aussi automatisée que possible pour l'obtention de l'espérance de vie des clients. Parce que le module IML de SAS n'était pas, lui non plus, disponible sur les postes de travail de la compagnie au moment de la réalisation de ce mémoire, deux choix s'offraient à nous : implémenter les calculs matriciels en utilisant de façon créative les fonctionnalités disponibles dans SAS/Base ou effectuer ces calculs avec le logiciel R. La première alternative a été tentée. Bien que la précision des résultats semblait adéquate, le délai requis pour effectuer les calculs excédait largement ce qui était envisageable (la cause étant que nous devions effectuer des calculs matriciels avec des fonctions et des outils qui n'ont pas été conçus à cet effet). La seconde alternative était évidemment d'utiliser R. Toutefois, la transition des calculs d'un logiciel à un autre implique quelques désagréments du point de vue technique et c'est pourquoi la possibilité d'obtenir la même mesure que celle donnée par le modèle de McFarland-Spilerman mais par une approche par simulations semblait attrayante.

Chapitre 3

RÉGRESSION LOGISTIQUE

Dans le modèle par chaînes de Markov de McFarland-Spilerman et dans l'approche par simulations présentés au chapitre 2, les probabilités de transition doivent être estimées. Afin d'obtenir des valeurs personnalisées en fonction des caractéristiques des clients, nous avons choisi de modéliser ces valeurs par le biais de régressions logistiques multinomiales. Ce chapitre est consacré à ce type de régression. Dans un premier temps, nous introduisons les modèles de choix discrets. Nous présentons ensuite la régression logistique binomiale puis nous généralisons à la régression logistique multinomiale.

3.1. MODÈLES DE CHOIX DISCRETS

Dans le cas classique de régression linéaire, la variable à expliquer est continue. À l'opposé, les modèles de choix discrets, qui englobent une panoplie de méthodes, sont utilisés lorsque la variable à expliquer est discrète. Par exemple, dans nos modèles, la variable à expliquer est l'état auquel le client se trouvera au temps suivant. Afin de représenter mathématiquement de tels modèles, différentes quantités sont utilisées.

Définition 3.1.1. *La **variable-réponse** est représentée par le vecteur* $\mathbf{Y}$ *de dimension* n *(où* n *est le nombre d'observations) dont l'élément* i *correspond à la valeur de la variable-réponse pour la* i^e *observation.*

Dans notre modèle par chaînes de Markov, les valeurs possibles de la variable-réponse sont les quatre états : avoir à la fois des polices automobiles et

résidentielles (état 1), n'avoir que des polices automobiles (état 2), n'avoir que des polices résidentielles (état 3) ou ne plus avoir de polices (état 4).

Définition 3.1.2. *Les différentes **variables explicatives** sont représentées par une matrice, notée* $\mathbf{X}$*, de taille* $p \times n$ *où* n *est le nombre d'observations et* p *est le nombre de variables explicatives. Chacune des colonnes correspond à un sujet et chacun des* p *éléments de cette colonne représente la valeur de la* p^e *variable explicative.*

Notons que dans la régression logistique usuelle, les variables explicatives sont supposées non aléatoires.

Définition 3.1.3. *Il est supposé que les valeurs de la variable-réponse dans la population suivent un certain modèle statistique. Les écarts des valeurs dans la population (vraies valeurs) aux valeurs du modèle supposé sont les **termes d'erreurs**,* $\boldsymbol{\epsilon}$*, qui sont confinés dans un vecteur de taille* n*.*

Définition 3.1.4. *La **fonction de lien** est la fonction* h *qui relie les variables explicatives à la variable-réponse. Mathématiquement, la relation entre la fonction de lien, les variables explicatives et la variable-réponse est donnée par*

$$E[\mathbf{Y}|\mathbf{X}] = h(\mathbf{X}). \tag{3.1.1}$$

Puisque nous ne pouvons observer la totalité de la variabilité (car si nous le pouvions, nous pourrions effectuer des prédictions parfaites), nous devons effectuer des hypothèses sur chacun des termes d'erreurs, ϵ_i. Nous supposons ainsi que ces termes suivent une certaine densité $f_i : \epsilon_i \sim f_i(e_i)$.

Nous sommes intéressés à prédire la valeur de la variable-réponse $\mathbf{Y}$ en fonction des valeurs observées des variables explicatives $\mathbf{X}$. La probabilité d'avoir observé $\mathbf{Y}$ étant donné $\mathbf{X}$ ne dépend que de la partie aléatoire non observable, $\boldsymbol{\epsilon}$. Ainsi, pour le i^e individu, la probabilité d'avoir observé la valeur Y_i de la variable-réponse étant donné le vecteur des caractéristiques X_i correspond à la probabilité d'avoir observé ϵ_i tel que $h(X_i) + \epsilon_i = Y_i$:

$$P(Y_i = y_i|X_i) = P(\epsilon_i \text{ t.q. } h(X_i) + \epsilon_i = y_i). \tag{3.1.2}$$

En notant « $I(h(X_i) + \epsilon_i = y_i)$ » la fonction indicatrice qui prend la valeur 1 lorsque $h(X_i) + \epsilon_i = y_i$ et 0 autrement, la probabilité cherchée est obtenue en

intégrant cette indicatrice sur l'ensemble des valeurs possibles de ϵ_i :

$$P(Y_i = y_i|X_i) = \int I(h(X_i) + \epsilon_i = y_i) f_i(\epsilon_i) d\epsilon_i. \tag{3.1.3}$$

Cette intégrale, une fois évaluée, ne dépend plus que des variables explicatives, X_i. En fonction de la probabilité d'avoir observé les différentes valeurs possibles de Y_i, il nous est possible de prendre une décision.

3.2. RÉGRESSION LOGISTIQUE BINOMIALE

La régression logistique binomiale est le type de régression logistique utilisée lorsque la variable à expliquer ne peut prendre que deux valeurs, codées 1 (l'événement d'intérêt s'est réalisé) ou 0 (l'événement d'intérêt ne s'est pas réalisé). Cette méthode tire son nom du fait qu'elle suppose que les termes d'erreurs ϵ suivent une distribution logistique, c'est-à-dire que

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}, \quad x \in \mathbb{R}, \tag{3.2.1}$$

ou, de façon équivalente, que la fonction de répartition soit

$$F(x) = \frac{1}{1 + e^{-x}}, \quad x \in \mathbb{R}. \tag{3.2.2}$$

Grâce à cette hypothèse et si $h(\cdot)$ correspond à la fonction identité, il est possible d'évaluer $P(Y_i = 0|X_i)$:

$$\begin{aligned} P(Y_i = 0|X_i) &= \int I(h(X_i) + \epsilon_i = 0) f_i(\epsilon_i) d\epsilon_i \\ &= \int I(X_i^t\beta + \epsilon_i = 0) f_i(\epsilon_i) d\epsilon_i \\ &= \int I(\epsilon_i = -X_i^t\beta) f_i(\epsilon_i) d\epsilon_i \\ &= F(-X_i^t\beta) \\ &= \frac{1}{1 + e^{-X_i^t\beta}}. \end{aligned}$$

La probabilité $P(Y_i = 1|X_i)$ est alors :

$$P(Y_i = 1|X_i) = 1 - P(Y_i = 0|X_i) = 1 - \frac{1}{1 + e^{-X_i^t\beta}} = \frac{e^{X_i^t\beta}}{1 + e^{X_i^t\beta}}. \tag{3.2.3}$$

Notons que la codification des variables est arbitraire ; nous pourrions tout aussi bien assigner la réalisation de l'événement à la valeur 0 et sa non réalisation à la valeur 1. Puisque la probabilité de l'un est le complémentaire de l'autre, l'interprétation des résultats ne changerait pas.

3.2.1. Ajustement du modèle

Puisque chacun des éléments de $\mathbf{Y}$, Y_i, prend soit la valeur 0, soit la valeur 1, nous avons que

$$Y_i \sim \text{Bernoulli}(P(Y_i = 1|X_i)) \tag{3.2.4}$$

(Kutner *et al.*, 2004). Notons qu'étant donné que l'étiquetage des classes est arbitraire, il est possible d'utiliser $P(Y_i = 1|X_i)$ ou $P(Y_i = 0|X_i)$; il suffit d'être cohérent avec ce choix tout au long de l'application de la méthode. Ainsi, la fonction de densité de Y_i est

$$f_i(y_i) = P(Y_i = y_i|X_i)^{y_i} \left(1 - P(Y_i = y_i|X_i)\right)^{1-y_i}. \tag{3.2.5}$$

Puisque les Y_i sont indépendants (mais non identiquement distribués car $P(Y_1|X_1) \neq P(Y_2|X_2) \neq \cdots \neq P(Y_n|X_n)$), la vraisemblance est

$$l(\boldsymbol{\beta}) = f(\mathbf{y}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} P(Y_i = y_i|X_i)^{y_i} \left(1 - P(Y_i = y_i|X_i)\right)^{1-y_i}. \tag{3.2.6}$$

Il faut alors utiliser des méthodes numériques afin d'optimiser la vraisemblance, ce qui nous permet d'obtenir l'estimateur $\widehat{\boldsymbol{\beta}}$.

3.2.2. Rapports de cotes

Il est pratique d'exprimer un modèle de régression en termes du prédicteur linéaire $\mathbf{X}^t\boldsymbol{\beta}$ étant donné que c'est cette quantité qui peut être facilement calculée à partir des estimateurs et des variables disponibles. Il est possible d'obtenir une telle expression pour la régression logistique en inversant l'équation (3.2.3). En appliquant la transformation inverse de l'exponentielle, le logarithme, et en effectuant des manipulations algébriques élémentaires, nous

obtenons l'expression

$$\log\left(\frac{p}{1-p}\right) = \mathbf{X}^t\boldsymbol{\beta}, \tag{3.2.7}$$

où $p = P(\mathbf{Y}|\mathbf{X})$ afin d'alléger la notation. Il n'est toutefois pas pratique d'interpréter la partie gauche de l'équation (3.2.7), c'est pourquoi nous interprétons habituellement la régression logistique en termes de rapport de cotes.

Définition 3.2.1. *Soit un événement A se produisant avec probabilité* p*. La cote de A est la probabilité que cet événement se produise divisée par la probabilité qu'il ne se produise pas :*

$$c = \frac{p}{1-p}. \tag{3.2.8}$$

On interprète la cote d'un événement ainsi : « l'événement A a c fois plus de chances de se réaliser que de ne pas se réaliser ».

Définition 3.2.2. *Soit deux événements A et B se produisant avec probabilités* p_A *et* p_B*. Le rapport de cotes de A sur B est donné par*

$$r_{AB} = \frac{p_A/(1-p_A)}{p_B/(1-p_B)}. \tag{3.2.9}$$

On interprète le rapport de cotes des événements A et B ainsi : « l'événement A a r_{AB} fois plus de chances de se produire que l'événement B ». Dans le cas de la régression logistique, ce n'est pas deux événements que nous comparons. En effet, supposons que nous ayons p variables explicatives. Nous cherchons à calculer le rapport de cotes de la i^e des variables explicatives. Afin d'illustrer le calcul sous-jacent, nous développons le terme $\mathbf{X}^t\hat{\boldsymbol{\beta}}$ en

$$\mathbf{X}^t\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\boldsymbol{\beta}}_i X_i + \cdots + \hat{\beta}_p X_p. \tag{3.2.10}$$

En appliquant l'exponentielle des deux côtés de l'équation (3.2.7), la cote du modèle ajusté est

$$c_{\beta_i} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_i X_i + \cdots + \hat{\beta}_p X_p\right) \tag{3.2.11}$$

alors que la cote de ce modèle dans lequel nous incrémentons la variable i d'une unité est

$$c_{\hat{\beta}_i+1} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_i(X_i+1) + \cdots + \hat{\beta}_p X_p\right). \tag{3.2.12}$$

En effectuant le ratio des équations (3.2.12) et (3.2.11), nous obtenons le rapport de cotes lié à une augmentation d'une unité de la variable i dans une régression logistique :

$$r_{\hat{\beta}_i} = \frac{c_{\hat{\beta}_i+1}}{c_{\hat{\beta}_i}} = \exp\left(\hat{\beta}_1\right). \tag{3.2.13}$$

L'interprétation du rapport de cotes lié à la i^e variable explicative va ainsi : « une augmentation d'une unité de la variable i entraîne une augmentation de r_{β_i} des chances que l'événement se produise » dans le cas où la modalité « 1 » a été codée comme la réalisation de l'événement et la modalité « 0 » comme sa non-réalisation.

3.3. RÉGRESSION LOGISTIQUE MULTINOMIALE

La régression logistique multinomiale est la généralisation de la régression logistique binomiale au cas où la variable à expliquer prend plus de deux valeurs. Supposons qu'il y ait m telles valeurs possibles, qui sont étiquettées $1, 2, \ldots, m$. Il est possible de se représenter la régression logistique multinomiale comme un ensemble de $m-1$ régressions logistiques binomiales dans lesquelles chacune des valeurs résultantes possibles sont comparées à une valeur de référence, disons la m^e valeur. Les $m-1$ équations qui correspondent aux cotes de ces modélisations sont

$$\frac{P(Y_i = 1|\mathbf{X}_i)}{P(Y_i = m|\mathbf{X}_i)} = \exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_1\right), \tag{3.3.1}$$

$$\frac{P(Y_i = 2|\mathbf{X}_i)}{P(Y_i = m|\mathbf{X}_i)} = \exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_2\right), \tag{3.3.2}$$

$$\vdots$$

$$\frac{P(Y_i = m-1|\mathbf{X}_i)}{P(Y_i = m|\mathbf{X}_i)} = \exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_{m-1}\right). \tag{3.3.3}$$

Notons que la définition de cote est ici légèrement différente de celle retrouvée à la section précédente ; plutôt que de comparer la probabilité qu'un événement se produise à la probabilité qu'il ne se produise pas, nous comparons la probabilité qu'il se produise versus la probabilité que l'événement de référence se produise. Il est possible de multiplier chacune des équations (3.3.1)

à (3.3.3) par $P(Y_i = m|\mathbf{X}_i)$ et de solutionner les équations qui en résultent pour cette probabilité. Nous obtenons ainsi chacune des probabilités des valeurs possibles de la variable à expliquer :

$$P(Y_i = 1|\mathbf{X}_i) = \frac{\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_1\right)}{1 + \sum_{k=1}^{m-1}\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_k\right)}, \tag{3.3.4}$$

$$P(Y_i = 2|\mathbf{X}_i) = \frac{\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_2\right)}{1 + \sum_{k=1}^{m-1}\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_k\right)}, \tag{3.3.5}$$

$$\vdots$$

$$P(Y_i = m-1|\mathbf{X}_i) = \frac{\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_{m-1}\right)}{1 + \sum_{k=1}^{m-1}\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_k\right)}, \tag{3.3.6}$$

$$P(Y_i = m|\mathbf{X}_i) = \frac{1}{1 + \sum_{k=1}^{m-1}\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_k\right)}. \tag{3.3.7}$$

Il serait possible de définir la dernière équation comme

$$P(Y_i = m|\mathbf{X}_i) = \frac{\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_m\right)}{1 + \sum_{k=1}^{m-1}\exp\left(\mathbf{X}_i^t\boldsymbol{\beta}_k\right)}, \tag{3.3.8}$$

mais le modèle serait alors surparamétré. L'utilisation de la contrainte $\boldsymbol{\beta}_m = 0$ permet de retrouver la valeur que nous avons déduite à partir de la représentation du modèle comme un ensemble de $m-1$ régressions logistiques binomiales.

Si dans la régression logistique binomiale Y_i admet une distribution de Bernoulli, cette distribution est multinomiale dans la régression logistique multinomiale :

$$Y_i \sim \text{Multinomiale}\left(P(Y_i = 1|\mathbf{X}_i), P(Y_i = 2|\mathbf{X}_i), \ldots, P(Y_i = m|\mathbf{X}_i)\right). \tag{3.3.9}$$

Il faut toutefois être prudent : il s'agit d'une distribution multinomiale, mais dans la majorité des applications, une seule observation par individu est disponible. Il est possible d'effectuer un parallèle avec la régression logistique binomiale de la section précédente, où la variable-réponse ne pouvait prendre que deux valeurs, 0 et 1. Chaque observation Y_i suivait alors une distribution binomiale dont une seule observation était disponible. Cela nous permet de

constater que la régression logistique multinomiale n'est qu'une généralisation de la régression logistique binomiale.

Les équations (3.3.4) à (3.3.6) sont les éléments constitutifs de la vraisemblance du modèle de régression logistique multinomiale. En définissant la fonction indicatrice

$$I(Y_i = j) = \begin{cases} 1 & \text{si l'entité } i \text{ prend la valeur } j \\ 0 & \text{sinon} \end{cases} \tag{3.3.10}$$

il est possible d'exprimer la vraisemblance du modèle de régression logistique multinomiale par

$$\begin{aligned} l(\boldsymbol{\beta}) &= \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \prod_{j=1}^{m} P(Y_i = j|\mathbf{X}_i)^{I(Y_i=j)} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m} \left(\frac{\exp\left(\mathbf{X}_i^t \boldsymbol{\beta}_j\right)}{1 + \sum_{k=1}^{m} \exp\left(\mathbf{X}_i^t \boldsymbol{\beta}_k\right)} \right)^{I(Y_i=j)} \end{aligned} \tag{3.3.11}$$

avec la contrainte $\beta_m = 0$. Il faut maximiser la vraisemblance de l'équation (3.3.11) à l'aide de méthodes numériques afin d'obtenir des estimateurs des $m-1$ vecteurs $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{m-1}$. Notons que l'équation (3.3.11) n'est qu'une représentation concise du modèle ; en pratique, il est plus utile de conceptualiser le problème comme l'optimisation conjointe des $m-1$ équations que l'on retrouve dans les équations (3.3.1) à (3.3.3).

3.3.1. Rapports de cotes

L'interprétation des rapports de cotes dans la contexte multinomial est légèrement plus complexe que dans le cas binomial. En effet, parce qu'une variable explicative donnée admet $m-1$ estimateurs, il faut calculer $m-1$ rapports de cotes pour obtenir un portrait de l'effet de cette variable sur l'ensemble des valeurs possibles de la variable à expliquer. Par surcroît, chacun de ces rapports de cotes doit être interprété par rapport à la modalité de référence. Supposons par exemple qu'une variable à expliquer puisse prendre trois valeurs, étiquettées 1, 2 et 3. Supposons qu'il n'y ait qu'une seule variable explicative, l'âge. Choisissons la modalité 3 comme modalité de référence. Pour connaître l'effet

de l'âge sur la variable-réponse, il faudrait calculer les rapports de cotes

$$\begin{aligned} r_{\beta_{11}} &= \frac{P(Y_i = 1|X_i, \beta_{10}, \beta_{11}+1)/P(Y_i = 3|X_i, \beta_{10}, \beta_{11}+1)}{P(Y_i = 1|X_i, \beta_{10}, \beta_{11})/P(Y_i = 3|X_i, \beta_{10}, \beta_{11}+1)} \\ &= \frac{\exp(\beta_{10} + (\beta_{11}+1)X_i)}{\exp(\beta_{10} + \beta_{11}X_i)} \\ &= \exp(\beta_{11}) \end{aligned} \tag{3.3.12}$$

et

$$\begin{aligned} r_{\beta_{21}} &= \frac{P(Y_i = 2|X_i, \beta_{20}, \beta_{21}+1)/P(Y_i = 3|X_i, \beta_{20}, \beta_{21}+1)}{P(Y_i = 2|X_i, \beta_{20}, \beta_{21})/P(Y_i = 3|X_i, \beta_{20}, \beta_{21}+1)} \\ &= \frac{\exp(\beta_{20} + (\beta_{21}+1)X_i)}{\exp(\beta_{20} + \beta_{21}X_i)} \\ &= \exp(\beta_{21}). \end{aligned} \tag{3.3.13}$$

Ces rapports de cotes s'interprètent comme suit :

> L'augmentation de une unité de la variable âge entraîne une augmentation de $r_{\beta_{11}}$ de la probabilité de prendre la valeur 1 plutôt que la valeur 3 et une augmentation de $r_{\beta_{21}}$ des chances de prendre la valeur 2 plutôt que la valeur 3.

Par exemple, si $r_{\beta_{11}} < 1$ et $r_{\beta_{21}} < 1$, alors une augmentation de l'âge entraîne une diminution des chances de prendre la valeur 1 plutôt que la valeur 3 : les individus transitent donc plus à la modalité 3 et similairement pour la valeur 2 par rapport à la valeur 3. Notons que la situation devient plus complexe à interpréter lorsque les rapports de cotes ont des signes différents.

3.3.2. Hypothèses

Le principe fondamental ayant permis le développement de la régresion logistique multinomiale a été proposé initialement par Luce (1959). L'hypothèse qu'il a effectuée est qu'étant donné un ensemble d'alternatives possibles M de la variable à expliquer, le rapport des probabilités des alternatives k et l, $\frac{P(Y_i=k)}{P(Y_i=l)}$

est le même pour tout ensemble M d'alternatives qui contiennent les alternatives k et l. C'est justement cette hypothèse qui a permis de décomposer la régression multinomiale à m alternatives en un ensemble de $m-1$ régressions binomiales. Luce a démontré que cette hypothèse – nommée indépendance des alternatives sans importance – permet d'obtenir l'expression de la probabilité de la i^e alternative sous la forme

$$P(Y_i = j) = \frac{w_i}{\sum_{k=1}^{m} w_k}, \tag{3.3.14}$$

où les w_k sont des quantités positives quantifiant le bénéfice associé à chacune des alternatives. Cette formulation en termes de bénéfices est liée au fait que les modèles de choix discrets ont été développés dans des contextes appliqués, notamment en économie et en sociologie. McFadden (1973) propose quant à lui la quantification des bénéfices par une fonction linéaire et l'utilisation de l'exponentielle pour garantir que ces bénéfices soient positifs, obtenant ainsi une expression de la forme

$$P(Y_i = j) = \frac{\exp\left(\mathbf{X}_i^t \boldsymbol{\beta}_j\right)}{\sum_{k=1}^{m} \exp\left(\mathbf{X}_i^t \boldsymbol{\beta}_k\right)}. \tag{3.3.15}$$

McFadden (1973) donne également les informations pertinentes à l'estimation de la vraisemblance de son modèle par la méthode du maximum de vraisemblance. L'hypothèse de l'indépendance des alternatives sans importance a donc permis le développement de la régression logistique multinomiale. Mais quel est son impact en termes pratiques ? McFadden (1980) utilise l'exemple des autobus bleus et rouges afin d'illustrer un cas où cette hypothèse n'est pas satisfaite. Dans cet exemple, les individus choisissent entre l'automobile (A) et un autobus bleu (B) avec une probabilité égale, c'est-à-dire que $P(A) = P(B) = \frac{1}{2}$, de telle sorte que le rapport $P(A)/P(B) = 1$. Si un nouveau mode de transport, un autobus rouge (R), devient disponible, l'hypothèse de l'indépendance des alternatives sans importance suppose que $P(A)/P(B)$ doit rester constant... cela implique donc que $P(A) = P(B) = P(R) = \frac{1}{3}$. Mais en pratique, nous nous attendons à ce que les individus n'aient pas de préférence entre les autobus bleus et rouges et que les probabilités soient $P(A) = 0{,}5$ et

$P(B) = P(R) = 0{,}25$, mais alors $P(A)/P(B) = 2$ et l'hypothèse est violée. Notons que dans le cas de notre modèle par chaîne de Markov, la définition des états nous paraît conforme à l'hypothèse de l'indépendance des alternatives sans importance étant donné que nos quatre états couvrent l'ensemble des alternatives possibles. En d'autres termes, nous ne voyons pas de quelle façon l'ensemble des alternatives possibles M pourrait changer.

Train (2009) identifie une autre hypothèse de la régression logistique multinomiale. Il s'agit du fait que les comportements expliqués par la régression logistique multinomiale ne peuvent être dus qu'aux facteurs observés. Plus encore, des caractéristiques observées, il faut que l'effet d'une variation d'une caractéristique sur le comportement des clients soit le même pour tous les clients. Dans le cas de notre modèle à quatre états pour le temps de vie, cela implique notamment qu'une augmentation de 200$ de la prime d'un individu aura le même impact pour *tous* les individus ayant les mêmes autres caractéristiques observées. Cette hypothèse peut être un peu forte : en raison des différentes situations de vie des clients qui sont dues à des caractéristiques non observées (quantité d'argent libre dans leur compte bancaire, destination-vacances anticipée pour l'année suivante, etc.) il est possible que devant une augmentation donnée de prime, des individus se comportent de façon différente, ce que ne peut expliquer la régression logistique. Une façon possible de ne pas effectuer cette hypothèse serait d'inclure des effets aléatoires dans notre modèle, ce qui permettrait d'effectuer des prédictions propres à chaque client. Il faut alors plusieurs années d'observations par client pour effectuer la modélisation, ce qui réduit substantiellement la quantité d'individus se qualifiant pour la modélisation en plus d'exclure les nouveaux clients de l'étude.

3.3.3. Lien entre la régression logistique et la distribution de Gumbel

Dans notre présentation de la régression logistique, nous avons mentionné que les termes d'erreurs liés à une alternative donnée sont distribués selon une distribution logistique, ce qui nous permettait d'intégrer le terme $P(Y_i|\mathbf{X}_i)$ de

l'équation (3.1.3). En fait, c'est la différence entre l'alternative d'intérêt et l'alternative de référence qui suit une distribution logistique. Comme le remarquera le lecteur désirant approfondir la régression logistique multinomiale, la présentation de cette méthode dans la littérature fait souvent référence à la distribution de Gumbel (McFadden, 1974) ou encore à la distribution valeurs extrêmes généralisée. Tel que mentionné à la section précédente, les justifications économétriques et sociologiques de cette classe de modèles font appel à la notion de bénéfice lié à chacune des alternatives possibles, les individus optant pour l'alternative ayant le bénéfice le plus grand (un traitement rigoureux de la notion de bénéfice est effectué par Chipman, 1960). La régression logistique multinomiale peut en fait être obtenue en supposant que les termes d'erreurs de chacune des *alternatives* suivent une distribution de Gumbel, dont la densité est

$$f(x) = e^{-x}e^{-e^{-x}} \quad x \in \mathbb{R} \tag{3.3.16}$$

et la fonction de répartition est

$$F(x) = e^{-e^{-x}} \quad x \in \mathbb{R}. \tag{3.3.17}$$

En effet, en regardant la différence entre les bénéfices liés à deux alternatives, nous obtenons une fonction logistique étant donné que la convolution de deux Gumbel résulte en une logistique ; dans notre présentation, nous avons simplement omis cette étape. Le recours à la distribution valeurs extrêmes généralisée est simplement expliqué par le fait que la distribution Gumbel en est un cas particulier.

Dans ce chapitre, nous avons présenté dans un contexte général la régression logistique multinomiale. Au chapitre 5, cette méthode sera appliquée à notre modèle par chaînes de Markov afin d'estimer les probabilités de transition. Le chapitre qui suit présente quant à lui les mesures d'espérance de vie qui seront calculées en utilisant notre modèle markovien.

Chapitre 4

MESURES DE TEMPS DE VIE

Dans les chapitres qui précèdent, nous avons expliqué brièvement la théorie sous-jacente aux modèles par chaînes de Markov que nous utilisons et nous avons expliqué les méthodes permettant d'estimer les probabilités de transition. Dans ce chapitre, nous expliquons ce qu'est l'espérance de vie et nous détaillons comment obtenir cette quantité dans chacun des modèles par chaîne de Markov utilisé.

4.1. ESPÉRANCE DE VIE

Lorsque nous étudions le temps de vie, plusieurs quantités peuvent être d'intérêt, telles que l'espérance de vie, le temps de vie médian, etc. La mesure de temps de vie qu'il nous a été demandé d'obtenir est l'espérance de vie. Nous commençons donc par définir ce qu'est l'espérance de vie et nous nous penchons sur son interprétation. Pour que cela soit possible, nous devons commencer par définir ce qu'est la fonction de survie.

Définition 4.1.1. *Soit* $T > 0$ *une variable aléatoire représentant le temps de vie qui admet une densité* f *et dont la fonction de répartition est* F*. Alors, la fonction de survie est donnée par*

$$S(t) = 1 - F(t) = 1 - P(T \leq t) = P(T > t). \tag{4.1.1}$$

La fonction de survie représente donc la probabilité qu'un client soit encore en vie au temps t. Le domaine qu'est l'analyse de survie propose, entre autres, différentes méthodes permettant d'estimer cette fonction de survie à

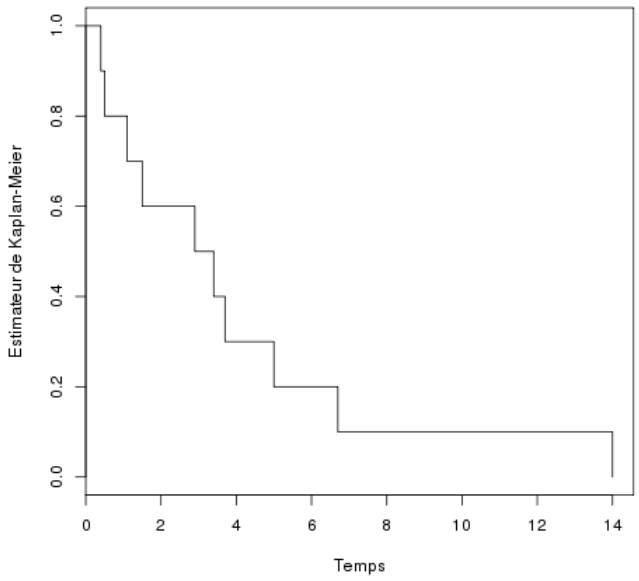
partir d'un échantillon. La plus connue est sans doute l'estimateur de Kaplan-Meier (Kaplan et Meyer, 1958). Nous utiliserons cet estimateur afin d'estimer la fonction de survie puisque cette dernière est une composante de l'espérance de vie.

Définition 4.1.2. *Soit les temps de vie observés dans un échantillon de taille* n, $t_1 \leq t_2 \leq \cdots \leq t_n$, *provenant d'une population dont la fonction de survie est* $S(t)$. *Supposons que* $t_i \leq t < t_{i+1}$. *Alors l'estimateur de Kaplan-Meier pour estimer* $S(t)$ *est donné par*

$$\hat{S}(t) = \prod_{t_i < t} \left(\frac{n_i - d_i}{n_i} \right), \tag{4.1.2}$$

où n_i *est le nombre d'individus en vie au temps* t_i *et* d_i *est le nombre d'individus décédés dans l'intervalle* $[t_i, t_{i+1}[$.

Exemple 4.1.1. *Supposons que nous avons 10 clients qui quittent la compagnie aux temps* 0,4; 0,5; 1,1; 1,5; 2,9; 3,4; 3,7; 5,0; 6,7; 14,0. *Alors l'estimateur de Kaplan-Meier est donné par*



Nous avons mentionné ci-dessus que la fonction de survie est utile afin de définir l'espérance de vie. Nous définissons maintenant l'espérance de vie au temps x, qui représente l'espérance du temps de vie restant au-delà du temps x.

Définition 4.1.3. *Soit* T *la variable aléatoire représentant le temps de vie dans une population et soit* $S(t)$ *sa fonction de survie. Alors l'espérance de vie au temps* t *est donnée par*

$$E[T|T > x] = \frac{\int_x^{+\infty}(t-x)f(t)dt}{S(x)} = \frac{\int_{T>t} S(t)dt}{S(x)}. \tag{4.1.3}$$

L'espérance de vie est définie à l'aide d'une intégrale. Il s'agit donc de l'aire sous la courbe de survie à droite de x. Pour un client donné, l'espérance de vie est ainsi une moyenne pondérée du temps qu'il est susceptible d'être un client de la compagnie, où les poids correspondent à sa probabilité d'être actif à chacun des temps. Afin d'estimer l'espérance de vie, il nous suffit de calculer l'aire sous la courbe de survie estimée par l'estimateur de Kaplan-Meier, c'est-à-dire qu'il suffit de substituer $\hat{S}(t)$ à $S(t)$ dans l'équation (4.1.3). Bien que la définition 4.1.3 soit valide pour tout x supérieur à 0, seul $\mu = E[T|T > 0] = E[T]$ est d'intérêt dans ce mémoire étant donné que le modèle sera ajusté sur l'année qui précède immédiatement celle d'intérêt et que nous voulons obtenir un estimateur de l'ensemble du temps de vie restant au client.

Étant donné que nos modèles discrétisent le temps, l'intégrale de l'équation (4.1.3) se transforme en somme. Puisque nous prédisons par intervalles d'un an, notre estimateur de l'espérance de vie devient

$$\hat{E}[T] = \sum_{i=0}^{k-1} \hat{S}(i), \tag{4.1.4}$$

où $\hat{S}(0)$, la probabilité d'être en vie au temps initial, vaut nécessairement 1 et où k est le nombre d'années après lesquelles l'espérance de vie est tronquée (nous supposons pour l'instant que nous effectuerons des prédictions jusqu'à un temps k assez élevé pour que $\hat{S}(k)$ soit près de zéro). Toutefois, étant donné que nous sommes dans un contexte de valeurs discrètes, nous avons apporté une correction à l'estimateur de Kaplan-Meier. En effet, l'estimateur de Kaplan-Meier est bon lorsque les temps observés prennent des valeurs continues et lorsque le nombre d'observations est assez grand. Puisque nous n'observons et ne prédisons les valeurs qu'à la fin d'une année, nous surestimerions l'espérance de vie si nous considérions l'estimateur de Kaplan-Meier sans correction

étant donné que nous savons que les individus ne sont pas décédés à la fin de l'année exactement, mais bien pendant l'ensemble de l'année. Un meilleur estimateur est donné par la moitié de la somme de l'aire supérieure reliant deux points (estimateur de Kaplan-Meier) et de l'aire inférieure reliant ces mêmes points (estimateur de Kaplan-Meier calculé avec une année de décalage), tel que représenté à la figure 4.1. Mathématiquement, la correction implique que l'aire sous la courbe est estimée par

$$\hat{E}[T] = \sum_{i=0}^{k-1} \hat{S}(i) + \frac{\hat{S}(k) - 1}{2} \tag{4.1.5}$$

plutôt que par (4.1.4), où $\frac{\hat{S}(k)-1}{2}$ est le facteur de correction qui différencie les deux équations. En fait, cette façon de faire est tout simplement la méthode d'analyse numérique servant à calculer une intégrale qu'est la méthode des trapèzes (Cohen, 2011). Cette méthode consiste à décomposer une fonction en intervalles et à approximer chacun de ces intervalles par une fonction linéaire. En d'autres termes, chacun des intervalles $[a, b]$ de la fonction f sont estimés par

$$\int_a^b f(x)dx \approx (b-a)\frac{f(a)+f(b)}{2}. \tag{4.1.6}$$

Dans notre cas, la fonction f est simplement la fonction d'espérance de vie de l'équation (4.1.3) et les intervalles sont $[0, 1], [1, 2], [2, 3], \ldots$ Nous effectuons donc l'hypothèse que la fonction d'espérance de vie est linéaire par morceaux et nous l'approximons en utilisant les points d'observation disponibles aux temps $0, 1, 2, 3, \ldots$ qui sont fournis par notre modèle prédictif.

Afin d'obtenir un estimateur approprié de l'espérance de vie, une étape additionnelle est requise. Il importe de réaliser que la fonction de survie est obtenue par une projection de la matrice de transition obtenue à partir des modèles markoviens pour des temps futurs. Ainsi, une hypothèse implicite de l'équation (4.1.4) (et donc de l'équation (4.1.5)) est que la fonction de survie estimée atteint la valeur zéro pour des temps assez grands. Or, nous nous sommes aperçus que ce n'était pas le cas. Afin de corriger l'estimateur, nous avons utilisé une fonction d'ajustement, ce qui est détaillé à la section 4.2. Bien que cet

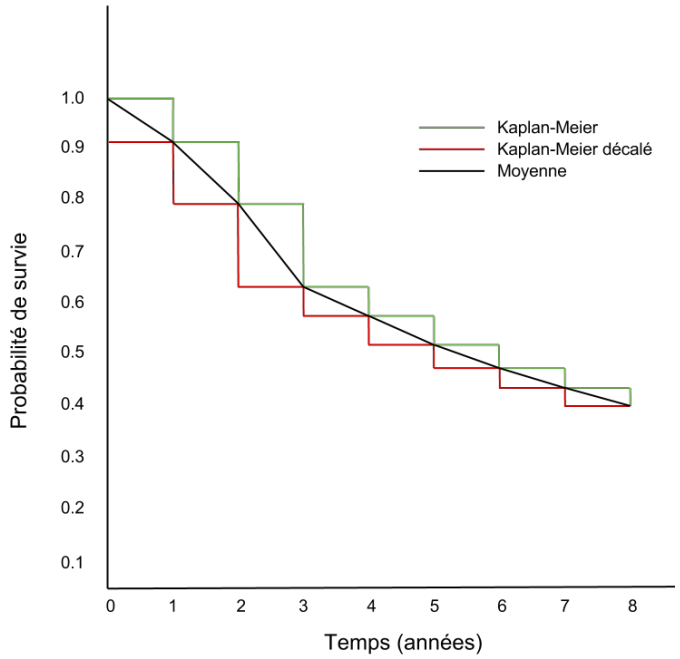


FIG. 4.1: Correction de l'estimateur de Kaplan-Meier pour temps de vie discrets

ajustement ait amélioré substantiellement la situation, un faible pourcentage des clients demeuraient actifs pour des temps de projections assez élevés (70 années). Nous avons alors décidé de cesser les projections. Ce faisant, les individus non décédés à la fin des projections (c'est-à-dire à la fin de l'« étude ») ne contribuaient pas au calcul de l'espérance de vie de telle sorte que notre estimateur était biaisé par le bas. Il s'agit en fait d'un cas de troncature à droite et nous avons corrigé notre estimateur pour cette situation.

Définition 4.1.4. *On dit qu'il y a* ***troncature à droite*** *lorsque que seuls les individus présentant certaines caractéristiques sont observés pendant une étude (Klein et Moeschberger, 1997).*

Afin de corriger l'estimateur de l'équation (4.1.5), il faut simplement diviser $\hat{E}[T]$ par la quantité d'individus non décédés à la fin de l'étude. Afin de faciliter la notation, les explications de cette assertion sont données pour le cas où la variable de temps est continue. Supposons que la variable de temps T continue admette une densité $f(t)$ et une fonction de répartition $F(t)$. Supposons qu'après un temps k, nous n'observons plus de temps de décès. La densité $f(t)$

ne somme plus à 1 : il faut donc la corriger. Dans cette situation, nous avons que

$$\int_{-\infty}^{+\infty} f(t)dt = \int_{-\infty}^{k} f(t)dt = 1 - S(K) \tag{4.1.7}$$

de telle sorte que

$$\int_{-\infty}^{k} \frac{f(t)}{1 - S(k)}dt = 1. \tag{4.1.8}$$

En posant $f(t,k) = \frac{f(t)}{1-S(k)}$ comme étant la densité corrigée pour la troncature à droite des observations au temps k, la densité somme maintenant à 1 :

$$\int_{-\infty}^{+\infty} f(t,k)dt = 1. \tag{4.1.9}$$

Afin de corriger l'espérance de vie, il suffit de remplacer $f(t)$ par $f(t,k)$ dans la définition de l'espérance de vie. En notant $E[T,k]$ l'espérance de vie corrigée pour la troncature à droite des observations au temps k, nous avons :

$$E[T,k] = \int_{-\infty}^{+\infty} t\, f(t,k)dt = \int_{-\infty}^{k} t\frac{f(t)}{1 - S(k)}dt \approx \frac{1}{1 - S(k)}E[T]. \tag{4.1.10}$$

Ainsi, nous avons que l'estimateur de l'équation (4.1.5) corrigé pour la troncature à droite au temps k est

$$\hat{E}[T,k] = \frac{1}{1 - S(k)}\left(\sum_{i=0}^{k-1} \hat{S}(i) + \frac{\hat{S}(k) - 1}{2}\right). \tag{4.1.11}$$

4.1.1. Calcul dans le modèle par chaîne de Markov simple

Dans le modèle par chaîne de Markov simple, nous utilisons les fréquences empiriques afin d'obtenir un estimateur de la matrice de transition P. Puis, en tirant profit de la propriété markovienne, nous obtenons les estimateurs subséquents en multipliant P avec elle-même un nombre de fois k afin d'obtenir des estimateurs des matrices de transition à k pas $\hat{P}^{(k)}_{\text{simple}}$. En prémultipliant cette matrice par le vecteur N_0 transposé contenant le nombre de clients à chacun des états au début de l'étude, nous obtenons la quantité de clients espérés à chacun des quatre états au temps k. En sommant le nombre de clients aux trois premiers états et en divisant par le nombre total de clients, nous obtenons une

estimation de la probabilité de survie au temps k :

$$\hat{S}_{\text{simple}}(k) = \frac{(1,1,1,0)N_0^t \hat{P}_{\text{simple}}^{(k)}}{\sum_{i=1}^{4} N_{0i}}. \tag{4.1.12}$$

4.1.2. Calcul dans le modèle de McFarland-Spilerman

Le calcul de l'espérance de vie dans le modèle basé sur la méthode de McFarland-Spilerman se fait de la même façon que pour le modèle par chaîne de Markov simple, à la différence près que les matrices de transition à k pas sont données par l'équation (2.5.2). Ainsi, nous avons que

$$\hat{S}_{\text{MS}}(k) = \frac{(1,1,1,0)N_0^t \hat{P}_{\text{MS}}^{(k)}}{\sum_{i=1}^{4} N_{0i}}. \tag{4.1.13}$$

4.1.3. Calcul dans l'approche par simulations

L'espérance de vie est obtenue d'une façon légèrement différente dans l'approche par simulations que dans les deux approches précédentes. Pour chacun des individus en vie au temps initial, nous appliquons l'algorithme suivant :

(1) obtenir les probabilités personnalisées de l'individu de se rendre aux états 1, 2, 3 et 4 (individus initialement à l'état 1), 1, 2 et 4 (individus initialement à l'état 2) ou 1, 3 et 4 (individus initialement à l'état 3) avec les régressions logistiques multinomiales ;

(2) générer une variable aléatoire uniforme entre 0 et 1 et prédire l'état suivant de l'individu en fonction de cette variable : pour un individu initialement à l'état 1, prédire que le prochain état sera l'état 1 si la variable de classification aléatoire est dans l'intervalle $[0, p_{11}]$, à l'état 2 si elle est dans l'intervalle $]p_{11}, p_{11} + p_{12}]$, à l'état 3 si elle est dans l'intervalle $]p_{11} + p_{12}, p_{11} + p_{12} + p_{13}]$ et à l'état 4 si elle est dans l'intervalle $]p_{11} + p_{12} + p_{13}, 1]$ (et similairement pour les clients se trouvant initialement aux états 2 et 3) ;

(3) si l'état prédit est le décès (état 4), arrêter l'algorithme. Sinon, incrémenter d'une année les variables explicatives qui dépendent du temps et recommencer à partir du point 1.

Lorsque tous les individus sont décédés ou lorsqu'est atteint un temps prédéterminé comme étant suffisamment grand, nous obtenons les probabilités de survie empiriques à chacun des temps. Nous sommes alors en mesure de représenter la courbe de survie et d'obtenir la valeur estimée de l'espérance de vie avec l'équation (4.1.5). Afin de réduire l'effet du classificateur aléatoire, nous répétons l'ensemble de l'expérience un nombre de fois assez élevé M ; nous pouvons ensuite obtenir l'espérance Monte-Carlo de ces M expériences. Afin d'obtenir une idée de la variabilité de chacune de ces simulations, nous pouvons calculer la variance Monte Carlo.

4.2. Modèle semi-paramétrique pour la fonction de survie

Lors de l'ajustement des régressions logistiques multinomiales, nous avons constaté que l'effet de la variable `client_since` (représentant l'ancienneté du client) qui découle d'une modélisation sur une année et utilisant l'ensemble des clients ne correspondait pas à l'effet réel de cette variable sur le temps de vie d'un client. En effet, globalement, nous avons trouvé (et cela nous semble en tous points logique) que les clients plus âgés sont en général plus fidèles que les clients plus jeunes. Ainsi, nous devrions en déduire, de façon unilatérale, que la probabilité de décès d'un client diminue avec le temps. Or, les clients étant avant tout des humains, l'événement de décès « ne plus être un client » est inéluctable étant donné qu'il surviendra ultimement en raison de causes naturelles telles que le décès physique des clients ou encore la vente de ses produits assurés (bon nombre de personnes âgées décident ultimement de ne plus conduire leur voiture ou encore quittent leur logis pour des foyers). Donc, après un temps assez grand, il serait souhaitable d'ajuster la fonction de survie afin de tenir compte de ce genre de facteurs. Afin de tenter de corriger la situation et d'obtenir une courbe de survie plus réaliste en son extrémité droite, nous avons décidé d'ajouter au modèle de Spilerman-McFarland une composante « non paramétrique » ; le modèle résultant est similaire au modèle semi-paramétrique de Cox pour la fonction de risque (Cox, 1972), à ceci près

que nous avons plutôt un modèle semi-paramétrique pour la fonction de survie. En effet, nous avons choisi de multiplier la courbe de survie qui découle du modèle markovien (et des régressions logistiques multinomiales qui en modélisent les probabilités de transition) par une courbe de survie représentant un effet plus réel de la variable `client_since`. Cette courbe de survie a donc un effet de ligne de base bornant supérieurement la courbe de survie prédite.

Se pose maintenant la question : comment obtenir une telle courbe à partir de nos données ? Nous avons exploité l'information contenue au temps 1, qui est, rapellons-le, le dernier point d'observation utilisé dans notre modélisation. À cet instant, nous disposons des valeurs de `client_since` de tous les individus dans l'échantillon. Certains de ces individus sont décédés, d'autres non. C'est donc une situation de censure à droite.

Définition 4.2.1. *On dit qu'il y a* ***censure à droite*** *au temps* k *lorsque nous savons que des individus décèdent à un temps plus grand que* k *sans connaître exactement le moment de leur décès (Klein et Moeschberger, 1997).*

En d'autres termes, nous savons que les individus sont encore en vie au temps k et nous pouvons les observer. C'est ce qui différencie la censure à droite de la troncature à droite, expliquée à la définition 4.1.4 : les individus censurés à droite font partie de l'étude alors que les individus tronqués à droite n'ont pu être observés. En tenant compte de cette censure, nous avons ajusté des densités sur la distribution de `client_since` en obtenant les estimateurs du maximum de vraisemblance des densités considérées et ce, dans l'optique d'utiliser la fonction de survie qui découle de cette densité ajustée. Évidemment, cette courbe n'est pas vraiment « non paramétrique » en ce sens que les paramètres des densités sont estimés. Nous qualifions malgré cela notre modèle de « semi-paramétrique » étant donné que l'estimation des paramètres impliqués a pour seul objectif d'obtenir une méthode purement objective pour la modélisation de notre espérance de vie. À l'instar du modèle semi-paramétrique de Cox, la fonction d'ajustement pourrait être sélectionnée sur des bases moins objectives et n'admettant pas de paramètres.

L'ensemble de ce qui a été mentionné à la section 4.1 sur le calcul de l'espérance de vie demeure valide ; nous utiliserons cependant dans nos calculs l'estimateur ajusté de la fonction de survie, que nous noterons $\hat{S}_{adj}(k)$, plutôt que la valeur $\hat{S}(k)$ obtenue directement du modèle.

4.3. Évaluation des mesures proposées

Maintenant que nous disposons de notre estimateur d'espérance de vie, nous discutons des moyens utilisés afin d'évaluer cette quantité. Le mandat qui nous a été assigné était d'obtenir l'espérance de vie des clients. Comme nous l'avons vu dans ce chapitre, l'espérance de vie dépend de l'ensemble des probabilités de décès futures des clients. Il nous est ainsi impossible d'obtenir une mesure d'espérance de vie empirique afin de valider nos modèles : il nous faudrait alors effectuer un suivi des clients que nous avons sélectionnés dans notre étude jusqu'à ce que chacun d'entre eux quittent la compagnie. Comme nous le verrons au chapitre 5, une fraction assez importante des clients sont très fidèles en assurance et cela fait en sorte qu'il sera irréaliste d'attendre que tous ces clients quittent la compagnie afin de valider notre mesure. Nous avons donc recours à des mesures alternatives permettant de valider partiellement notre modèle. Puisque nous avons conservé quatre années de données en plus de l'année utilisée pour la modélisation, nous pouvons effectuer des comparaisons des valeurs prédites et des valeurs empiriques tant que les quantités prédites sont limitées à ces cinq années.

4.3.1. Probabilités de survie prédites après cinq ans

La première mesure utilisée afin de valider nos modèles sont les probabilités de décès prédites après cinq ans. Afin d'obtenir ces quantités, nous utilisons nos modèles afin d'obtenir la probabilité de *chaque individu* d'être en vie après cinq années ; nous effectuons ensuite la moyenne au niveau de l'échantillon. Cette mesure nous est utile pour deux raisons. D'une part, elle nous permet

de comparer la probabilité de survie moyenne après cinq ans avec la probabilité empirique d'être en vie après cinq ans. D'autre part, nous pouvons calculer cette métrique autant pour ceux qui sont en réalité décédés après cinq ans que pour ceux qui sont réellement en vie après cinq ans. Nous pouvons ainsi déterminer si le modèle permet de discriminer les bons individus de même que le degré avec lequel il permet cette discrimination. Si nos modèles étaient parfaits, nous aurions une probabilité de survie prédite de 0% chez les individus qui sont réellement décédés et une probabilité de survie prédite de 100% chez les individus qui sont réellement en vie. Puisque nous effectuons des prédictions individuelles à partir de tendances observées à partir de l'échantillon, que ces comportements individuels sont difficiles à capter par un modèle aussi général et que les variables disponibles sont somme toute assez limitées, nous ne nous attendons pas à avoir une discrimination très grande. Le seul fait de discerner des différences sera donc satisfaisant.

4.3.2. Espérance de vie censurée après cinq ans

La deuxième mesure utilisée afin de valider nos modèles est l'espérance de vie censurée après cinq ans. Il s'agit simplement de l'espérance de vie dans laquelle nous supposons que la probabilité de survie est nulle au-delà de cinq années. Notons qu'il ne s'agit pas d'une espérance de vie *tronquée* après cinq ans étant donné que les individus qui ne sont pas décédés après cette période contribuent au calcul ; nous supposons toutefois que ces clients décèdent immédiatement après ces cinq années de vie. Afin de la calculer, nous utilisons l'équation (4.1.4) en imposant la contrainte

$$S(k) = 0 \quad \forall k \geq 6. \tag{4.3.1}$$

Nous estimons alors cette quantité par 0 dans (4.1.4) pour les temps appropriés. L'obtention d'une métrique ayant une interprétation sur cinq années était également pertinente pour la compagnie. En effet, les projets d'entreprise sont souvent évalués selon leurs bénéfices sur cinq ans.

Dans ce chapitre, nous avons présenté les différents estimateurs de l'espérance de vie utilisés, le modèle semi-paramétrique pour la fonction de survie ainsi que les mesures d'évaluation de nos modèles. Dans le chapitre qui suit, nous appliquons l'ensemble des concepts présentés dans les quatre premiers chapitres aux données disponibles.

Chapitre 5

RÉSULTATS

Ce chapitre présente les performances des différents modèles qui ont étés présentés dans les chapitres précédents. Nous commençons avec une description des données disponibles, pour ensuite présenter les performances comparatives des trois approches proposées.

5.1. ANALYSE DESCRIPTIVE DES DONNÉES

Le jeu de données contient 216 305 clients qui ont chacun six points d'observation pendant la période d'étude. La période d'étude s'étend du 1er janvier 2006 au 1er janvier 2011 et chaque 1er janvier représente un point d'observation des clients. Notons que le jeu de données a été créé à partir de différentes sources de données (information sur les comptes-clients, information sur les polices automobiles, information sur les polices résidentielles, information sociodémographique) et qu'un traitement élaboré des différentes sources de données primaires a dû être effectué afin de disposer de l'information dans un format adéquat à nos analyses. Un total de 24 variables ont été considérées pour les modèles (se référer au tableau 5.1 pour la liste exhaustive de ces variables, au tableau 5.2 pour les moyennes des variables continues et aux tableaux 5.3 et 5.4 pour les fréquences des variables nominales ; notons que pour des raisons de confidentialité des numéros ont étés assignés aux groupes). De ces 24 variables, 8 contiennent de l'information sur les comptes-clients ou sur les polices et 16 contiennent de l'information sociodémographique, information agrégée par code postal et dérivée du recensement de Statistique Canada. Afin de ne

retenir que les variables les plus pertinentes, une sélection de variables de type pas-à-pas a été effectuée lors de la modélisation des probabilités de transition utilisant les régressions logistiques multinomiales. L'ensemble des manipulations, de l'extraction aux prédictions – hormis la modélisation du modèle de McFarland-Spilerman , qui utilise le logiciel R version 2.10 – ont été réalisées à l'aide du logiciel SAS. Le serveur utilisé pour l'extraction des données disposait de la version 8.2 de SAS alors que celui utilisé pour la modélisation utilisait la version 9.2.

Les modèles qui ont été ajustés reposent sur une discrétisation du temps. Néanmoins, il est intéressant de regarder des statistiques descriptives de la variable représentant le temps de vie. Commençons par noter qu'après 5 ans, 32,30% des clients étaient « décédés ». Pour les clients qui sont décédés, un histogramme du temps de décès se retrouve à la figure 5.1. Nous remarquons que le nombre de décès diminue lorsque le temps augmente. Ce phénomène est relativement intuitif : plus un client reste longtemps avec sa compagnie, plus élevée est sa probabilité de demeurer avec elle une année de plus. Dans un autre ordre d'idées, l'observation de la figure 5.1 permet de se rendre compte qu'il semble y avoir une certaine saisonnalité dans les temps de décès. Il est bien connu que certains mois sont plus propices à la vente de certains types d'assurance (par exemple, les gens assurent plus souvent des motos vers le mois d'avril que vers le mois de novembre)... la majorité des annulations prenant effet à la fin des contrats, il est tout aussi logique que ces tendances soit discernables dans les temps de décès. Compte tenu du fait que nous tentons de prédire la courbe de survie sur l'ensemble de ses valeurs futures, la présence d'une saisonnalité ne nous inquiète pas outre mesure. En effet, sur l'ensemble de la courbe de survie, l'effet de la tendance saisonnière ne sera que très limitée, d'autant plus qu'elle semble s'estomper pour les années plus grandes (voir années 4 et 5 sur la figure 5.1). Plus important encore, l'utilisation d'une modélisation impliquant une discrétisation du temps et qui utilise des intervalles d'un an protège nos estimateurs de l'effet de cette saisonnalité, le nombre de personnes décédant à l'intérieur d'une période d'un an ne dépendant pas de la

TAB. 5.1: Description des variables disponibles (CP = variables sociodémographiques agrégées par code postal)

Variable	Description
client_group	Le groupe d'affinité auquel appartient le client. Les groupes qui étaient trop petits pour être modélisés ont étés regroupés dans une catégorie « Autres ».
client_sex_owner1	Le sexe du titulaire principal du compte (F = Femme, M = Homme).
ind_sec_owner	Indicateur de la présence d'un titulaire secondaire.
tot_premium	Prime totale (somme de la prime de toutes les polices actives du compte).
client _age_owner1	Âge du titulaire principal du compte.
client_since_yr	Temps (en années) depuis l'ouverture du compte.
senior	Le titulaire principal est-il âgé de plus de 65 ans ?
new_client	Le client est-il actif depuis moins d'un an ?
avg_person_per_hh	Nombre moyen d'habitants par ménage (CP).
am_income_avg_total	Revenu moyen par ménage (CP).
credit_score_census	Cote de crédit moyenne (CP).
rt_total_popul_males	Pourcentage d'hommes (CP).
rt_ages_20_29	Pourcentage d'individus âgés entre 20 et 29 ans (CP).
rt_ages_60plus	Pourcentage d'individus âgés de plus de 60 ans (CP).
rt_tot_visible_minority	Pourcentage d'individus se définissant comme « minorité visible » (CP).
rt_tot_immigrant	Pourcentage d'immigrants (CP).
rt_univ_over_tot_pop_educ	Pourcentage d'individus scolarisés qui détiennent un baccalauréat (CP).
rt_employ_15	Taux d'emploi (chez les 15 ans et plus) (CP).
rt_children_0_17	Indice du nombre d'enfants par ménage (CP).
rt_family_legally_married	Pourcentage de familles mariées (CP).
rt_owned	Pourcentages d'habitations dont les occupants sont les propriétaires (CP).
rt_occ_sales_service	Pourcentage d'individus travaillant dans le domaine de la vente et des services (CP).
rt_blue_collar	Pourcentage de cols bleus (CP).
rt_white_collar	Pourcentage de cols blancs (CP).

saisonnalité mensuelle. Notons qu'une modélisation mensuelle plutôt qu'annuelle serait quant à elle influencée par cette saisonnalité ; advenant le cas où nous souhaiterions raffiner le modèle développé dans ce mémoire en ce sens, il serait intéressant d'incorporer une tendance saisonnière à notre modélisation.

TAB. 5.2: Statistiques descriptives des variables continues au début de l'étude

Variable	Moyenne	Écart-type	Minimum	Maximum
tot_premium	2 102,41	1 190,02	100,00	17 309,00
client_age_owner1	41,98	11,91	16,00	98,90
client_since_yr	5,97	6,27	0,00	64,17
avg_person_per_hh	2,88	0,64	1,00	5,30
am_income_avg_total	42 010,42	23 437,83	8 737,00	531 658,00
credit_score_census	747,85	37,35	0,00	870,17
rt_total_popul_males	0,49	0,03	0,00	1,00
rt_ages_20_29	0,14	0,07	0,00	1,00
rt_ages_60plus	0,17	0,10	0,00	1,00
rt_tot_visible_minority	0,34	0,29	0,00	1,00
rt_tot_immigrant	0,38	0,21	0,00	1,00
rt_univ_over_tot_pop_educ	0,34	0,16	0,00	1,00
rt_employ_15	0,64	0,10	0,00	1,00
rt_children_0_17	0,79	0,28	0,00	3,00
rt_family_legally_married	0,53	0,12	0,00	1,00
rt_owned	0,74	0,29	0,00	1,00
rt_occ_sales_service	0,22	0,08	0,00	1,00
rt_blue_collar	0,18	0,13	0,00	1,00
rt_white_collar	0,60	0,16	0,00	1,00

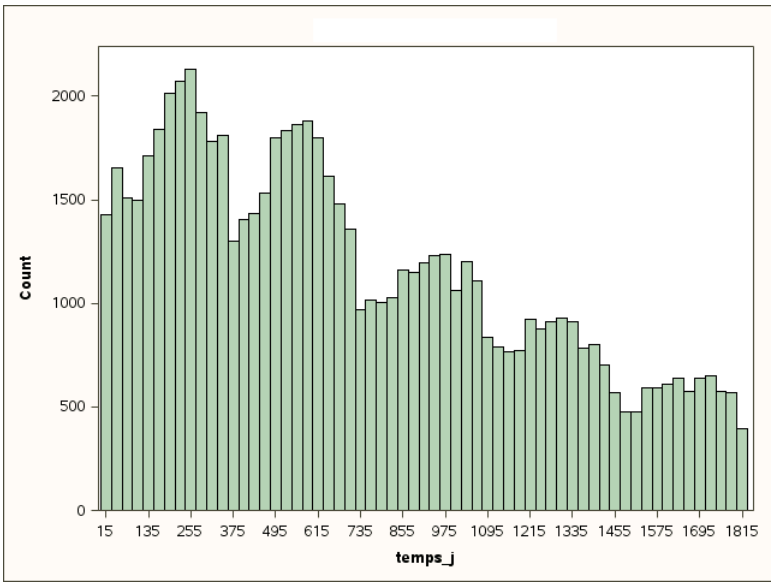


FIG. 5.1: Histogramme du temps de décès (en jours) des individus qui sont décédés

Pour terminer cette brève analyse descriptive des données, nous nous intéressons aux différents « statuts » que peuvent prendre les clients puisqu'ils

TAB. 5.3: Fréquences et fréquences relatives de la variable explicative client_group

Groupe	Fréquence	Fréquence relative (%)
1	7 164	3,31
2	33 537	15,50
3	2 580	1,19
4	7 183	3,32
5	6 958	3,22
6	5 743	2,66
7	10 912	5,04
8	3 002	1,39
9	6 087	2,81
10	8 179	3,78
11	2 436	1,13
12	3 879	1,79
13	2 338	1,08
14	3 889	1,80
15	6 904	3,19
16	2 570	1,19
17	3 450	1,59
18	4 146	1,92
19	18 341	8,48
20	13 399	6,19
21	4 903	2,27
22	2 719	1,26
23	6 897	3,19
24	11 017	5,09
Autres	38 072	17,60

TAB. 5.4: Fréquences relatives de la variable client_sex_owner1

client_sex_owner1	Fréquence relative (%)
F	36,37
H	63,63

sont d'une importance primordiale dans les modèles qui ont étés ajustés. En effet, ces statuts constituent les états des modèles par chaînes de Markov. Rappelons qu'un client peut avoir à la fois des polices automobiles et résidentielles (état 1), n'avoir que des polices automobiles (état 2), n'avoir que des polices résidentielles (état 3) ou encore ne plus avoir aucune police (« décès », état 4). Le tableau 5.5 présente l'évolution du nombre de clients à chacun des états à

TAB. 5.5: Évolution des états des clients à travers le temps

État	Temps 0	Temps 1	Temps 2	Temps 3	Temps 4	Temps 5
1	118 406	119 832	117 842	114 531	112 011	109 887
(%)	(54,74)	(55,40)	(54,48)	(52,95)	(51,78)	(50,80)
2	77 303	57 248	41 650	32 755	26 692	23 243
(%)	(35,74)	(26,47)	(19,26)	(15,14)	(12,34)	(10,75)
3	20 596	17 980	16 545	15 446	14 355	13 236
(%)	(9,52)	(8,31)	(7,65)	(7,14)	(6,64)	(6,12)
4	0	21 245	40 268	53 573	63 247	69 939
(%)	(0,00)	(9,82)	(18,62)	(24,77)	(29,24)	(32,33)
Total	216 305	216 305	216 305	216 305	216 305	216 305
(%)	(100,00)	(100,00)	(100,00)	(100,00)	(100,00)	(100,00)

travers le temps. Le temps 0 est le début de l'étude et le temps 5 est le dernier temps d'observation. Les valeurs aux temps 1, 2, 3 et 4 représentent le nombre de clients qui se trouvaient aux différents états exactement une année après le début de l'étude, deux années après le début, etc. Quelques constats peuvent être effectués en observant le tableau 5.5. Premièrement, au temps initial, aucun individu n'est décédé : cela découle du fait que nous avons sélectionné les clients qui étaient « en vie » (actifs) à un moment donné et nous avons effectué le suivi de leurs produits pendant les cinq années qui suivent. Nous constatons également que la proportion de clients à l'état 1 varie moins entre le temps 0 et le temps 5 (de 54,74% à 50,80%, diminution de 7,20%) que les proportions aux états 2 (de 35,74% à 10,75%, diminution de 69,92%) et 3 (de 9,52% à 6,12%, diminution de 35,71%). Cela illustre un autre phénomène connu en assurance de biens : les clients qui ont plus d'un type de produits assurés (résidentiel et automobile) sont plus fidèles à leur assureur que ceux qui n'ont qu'un des deux produits ; c'est ce qui explique l'omniprésence de rabais multi-produits au sein des compagnies d'assurance.

Un autre présupposé souvent effectué dans la modélisation de données en assurance est que les nouveaux clients se comportent différemment des anciens clients. La durée d'une police d'assurance étant d'une année, nous définissons habituellement « nouveau client » comme un client étant actif depuis moins d'un an, le premier renouvellement tenant lieu de rite de passage. Notre

modélisation comporte une variable indiquant si les clients sont actifs depuis moins d'un an, qui est dérivée de la variable *client_since_yr*. Afin de visualiser les différences potentielles de comportements entre les anciens et les nouveaux clients, nous avons représenté à la figure 5.2 les courbes de survie empiriques pour les nouveaux clients, pour les anciens clients et la courbe de survie globale. Nous constatons, en effet, que les nouveaux clients semblent décéder à une fréquence plus élevée que les anciens clients pendant les premières années de l'étude. En regardant les distributions des états initiaux auprès de ces deux strates de clients (voir tableau 5.6), nous constatons que les anciens clients semblent beaucoup plus souvent se situer à l'état 1 (produits résidentiels et automobiles assurés) que les nouveaux clients, ce qui est cohérent avec ce que nous avons mentionné précedemment quant au fait que les clients à l'état 1 devraient être plus fidèles.

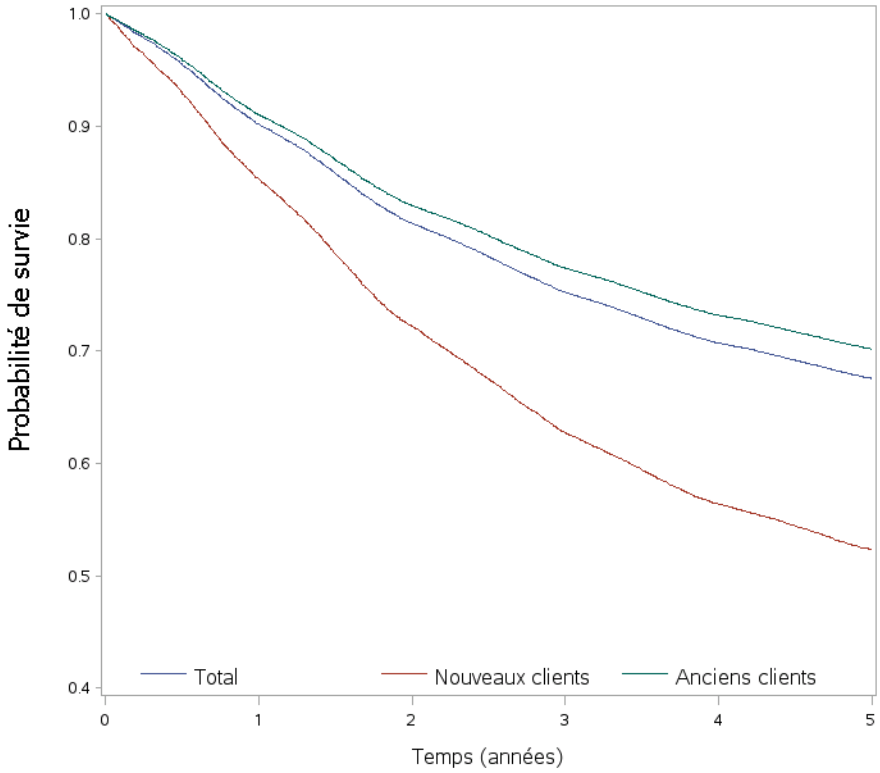


FIG. 5.2: Courbes de survie empiriques

TAB. 5.6: Distribution des états initiaux

État	Nouveaux clients (n = 32 114)	Anciens clients (n = 184 191)	Tous les clients (n = 216 305)
1	31,50%	58,79%	54,74%
2	59,62%	31,57%	35,74%
3	8,88%	9,63%	9,52%
4	0,00%	0,00%	0,00%

L'ensemble des statistiques descriptives qui ont été présentées dans cette section ont été réalisées pour les 6 points d'observation des clients et sur l'ensemble des clients disponibles. Dans les modélisations qui ont été choisies, nous n'utilisons que la première année disponible (passage du temps 0 au temps 1) afin d'obtenir les espérances de vie prédites. Par ailleurs, nous avons partitionné les données disponibles en deux sous-ensembles : un ensemble d'entraînement (70%) sur lequel les régressions ont été ajustées, et un ensemble de validation (30%), qui nous servait à s'assurer que la qualité des mesures obtenues n'était pas dépendante du jeu de données et à constater l'étendue des disparités des mesures prédites à l'aide d'un jeu de données indépendant. Les sections qui suivent présentent donc les résultats à la fois sur le jeu de données d'entraînement et sur le jeu de données de validation. Dans l'ajustement des modèles, les années restantes (points d'observation qui suivent le temps 1) n'ont pas été considérés ; nous les avons réservées pour nos mesures de validation de la section 5.6.

5.2. MODÈLE PAR CHAÎNES DE MARKOV SIMPLE

Dans le modèle par chaînes de Markov simple, nous utilisons les fréquences relatives de transition entre les états comme estimateurs des probabilités de transition. En utilisant les deux premiers points d'observation des clients, nous

obtenons l'estimateur de la matrice de transition à un pas $\hat{P}_{\text{simple}}$ suivante :

$$\hat{P}_{\text{simple}} = \begin{pmatrix} 0{,}9164 & 0{,}0200 & 0{,}0171 & 0{,}0464 \\ 0{,}1223 & 0{,}7098 & 0{,}0000 & 0{,}1678 \\ 0{,}0937 & 0{,}0000 & 0{,}7682 & 0{,}1381 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}. \tag{5.2.1}$$

Notons que les valeurs nulles correspondant respectivement à $\hat{p}_{23}$ et $\hat{p}_{32}$ n'ont pas été estimées ; nous leur avons assigné ces valeurs étant donné que nous avons déterminé que cette transition survenait tellement rarement que leur modélisation n'était pas d'intérêt. Il est possible de diagonaliser la matrice $\hat{P}_{\text{simple}}$ afin d'obtenir une expression facile à calculer de l'estimateur de la matrice de transition à k pas :

$$\begin{aligned} \hat{P}^{k}_{\text{simple}} = & \begin{pmatrix} 0{,}5000 & -0{,}7807 & -0{,}0795 & -0{,}1017 \\ 0{,}5000 & -0{,}4265 & -0{,}2301 & 0{,}9814 \\ 0{,}5000 & -0{,}4568 & 0{,}9699 & 0{,}1631 \\ 0{,}5000 & 0{,}0000 & 0{,}0000 & 0{,}0000 \end{pmatrix} \\ & \times \begin{pmatrix} 1{,}0000 & 0{,}0000 & 0{,}0000 & 0{,}0000 \\ 0{,}0000 & 0{,}9366 & 0{,}0000 & 0{,}0000 \\ 0{,}0000 & 0{,}0000 & 0{,}7573 & 0{,}0000 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 0{,}7031 \end{pmatrix} \\ & \times \begin{pmatrix} 0{,}0000 & 0{,}0000 & 0{,}0000 & 2{,}0000 \\ -1{,}1569 & -0{,}1001 & -0{,}1185 & 1{,}3755 \\ -0{,}4428 & -0{,}2032 & 0{,}9465 & -0{,}3005 \\ -0{,}6065 & 0{,}9278 & 0{,}1704 & -0{,}4917 \end{pmatrix}. \end{aligned} \tag{5.2.2}$$

À partir de cette expression, nous pouvons obtenir la courbe de survie prédite (voir figure 5.3), dont l'aire sous la courbe donne un estimateur de l'espérance de vie de la population. Cette aire, donc l'espérance de vie, vaut 13,38 années.

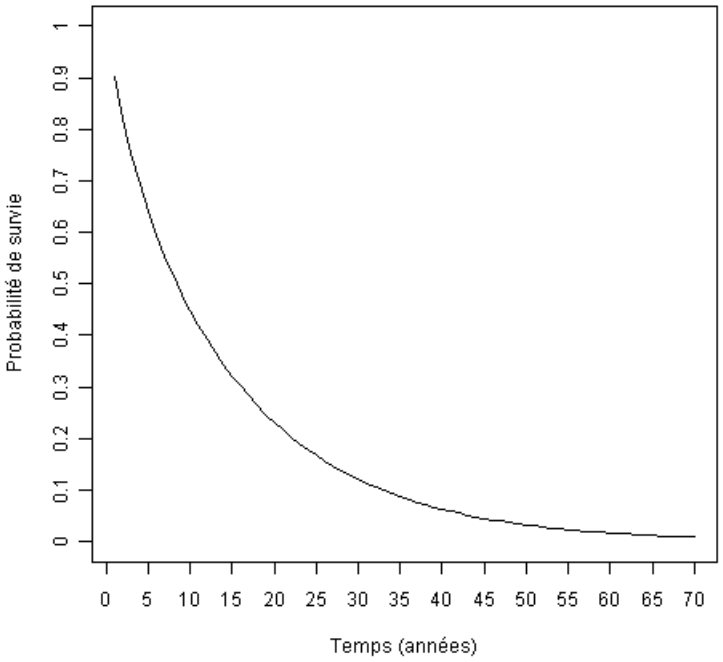


FIG. 5.3: Courbe de survie prédite, modèle de Markov simple

5.3. MODÈLE DE MCFARLAND-SPILERMAN

Dans cette section, nous présentons l'ensemble des résultats reliés au modèle de McFarland-Spilerman : les probabilités de transition par les régressions logistiques multinomiales, l'obtention des estimateurs des matrices de transition à k pas, la sélection de la fonction d'ajustement et les valeurs obtenues des estimateurs de l'espérance de vie.

5.3.1. Probabilités de transitions

Afin d'obtenir l'espérance de vie dans la méthode de McFarland-Spilerman avec fonction d'ajustement, il faut d'abord modéliser les probabilités de transition entre les états. Rappelons que trois régressions logistiques multinomiales ont été ajustées afin de modéliser les probabilités de transition des clients qui se trouvent initialement aux états 1 (deux types de produits), 2 (produits automobiles seulement) et 3 (produits résidentiels seulement). Les tableaux 5.7 et 5.8

présentent l'ensemble des résultats des trois régressions logistiques multinomiales par le biais des rapports de cotes résultant de ces régressions. Les deux premières colonnes de ces tableaux contiennent les variables des régressions logistiques (le second tableau est dédié à la variable *client_group*) et l'unité par rapport à laquelle les rapports de cotes doivent être interprétés. Les colonnes qui suivent sont regroupées en trois catégories en fonction des régressions représentées par les rapports de cotes (régressions pour les individus initialement à l'état 1, à l'état 2 et à l'état 3). Afin de comprendre la signification des rapports de cotes présentés, considérons les états a, b et c et la variable explicative X. L'interprétation du rapport de cotes dans un contexte de régression logistique multinomiale va ainsi (en supposant que la catégorie de référence pour la modélisation soit l'état c) :

> Un client donné a RC_{ab} fois plus de chances d'aller de l'état a à l'état c que de l'état b à l'état c lorsque la variable explicative augmente d'une unité.

Dans le cas où X est une variable explicative catégorielle prenant les modalités E et F, l'interprétation du rapport de cotes attribuable à la variable X (E vs F), c'est-à-dire que la catégorie E est contrastée à la catégorie F, devient :

> Un client donné a RC_{ab} fois plus de chances d'aller de l'état a à l'état c que de l'état b à l'état c lorsqu'il possède la valeur E de la variable explicative que lorsqu'il possède la valeur F.

Puisque les trois régressions sont ajustées indépendamment les unes des autres, les variables retenues par la sélection de variables de type pas-à-pas ne sont pas toutes les mêmes ; c'est ce qui explique l'absence de certaines valeurs dans le tableau 5.7. Les astérisques indiquent si les rapports de cotes sont significatifs au seuil de signification de 5%.

5.3.1.1. *Remarque sur la signification et la significativité des rapports de cotes*

Nous sommes prudents quant au sens à donner aux rapports de cotes des variables impliquées dans les régressions logistiques multinomiales. Nous

TAB. 5.7: Rapports de cotes des régressions logistiques multinomiales (modèle de McFarland-Spilerman)

Variables	Unités	De l'état 1			De l'état 2		De l'état 3	
		RC_{14}	RC_{24}	RC_{34}	RC_{14}	RC_{24}	RC_{14}	RC_{34}
tot_premium	milliers	*1,113	*0,866	*0,820	*1,250	*1,124	*5,340	*3,924
senior	0 vs 1	*1,453	0,835	0,834	*1,361	*1,323	1,160	1,007
new_client	0 vs 1	1,087	0,847	*1,548	*0,874	*1,068	*0,548	*1,184
sex_owner1	F vs M	*1,120	*1,212	*1,154	1,053	*1,111	1,131	*1,216
ind_sec_owner	0 vs 1	-	-	-	*0,672	1,027	*0,699	1,118
client_age_owner1	années	*1,040	1,002	*1,015	*1,009	*1,028	1,003	*1,029
avg_person_per_hh	personnes	*1,238	*1,348	*1,822	0,992	*1,182	0,993	*0,841
am_income_avg_total	milliers	0,999	1,002	1,003	0,999	1,001	-	-
credit_score_census	centaine de points	*1,099	0,892	0,952	*1,109	*1,334	1,174	*1,183
rt_ages_20_29	pourcent	*0,990	*0,987	*0,982	1,000	*0,995	*0,981	*0,987
rt_children_0_17	pourcent	0,999	*0,994	*0,994	-	-	-	-
rt_tot_visible_minority	pourcent	*0,994	*0,996	1,000	*0,997	*0,996	-	-
rt_tot_immigrant	pourcent	-	-	-	1.001	*0,997	-	-
rt_family_legally_married	pourcent	0,998	0,991	*0,970	-	-	-	-
rt_owned	pourcent	*1,005	*0,996	*1,006	-	-	-	-
rt_employ	pourcent	-	-	-	*1,005	0,999	-	-
rt_white_collars	pourcent	*1,004	0,995	1,005	-	-	*0,993	1,001
rt_univ_over_tot_pop_educ	pourcent	-	-	-	-	-	-	-
rt_tot_popul_males	pourcent	-	-	-	-	-	1,001	*1,020

Les rapports de cotes marqués d'un astérisque (*) sont significatifs au seuil de signification de 5%.

Des valeurs sont manquantes (-) lorsque les variables n'ont pas été retenues dans la sélection de variables.

TAB. 5.8: Rapports de cotes des régressions logistiques multinomiales (modèle de McFarland-Spilerman) pour la variable client_group

Variables	Unités	De l'état 1			De l'état 2		De l'état 3	
		RC_{14}	RC_{24}	RC_{34}	RC_{14}	RC_{24}	RC_{14}	RC_{34}
client_group	1 vs Others	*0,214	*0,591	0,937	*0,180	*0,381	*0,113	*0,688
client_group	2 vs Others	*2,188	1,250	1,148	*2,022	*1,807	*1,888	*1,524
client_group	3 vs Others	*0,376	*0,468*	1,321	0,780	*0,607	*0,333	0,937
client_group	4 vs Others	*1,230	1,000	*1,482	*1,779	*1,390	0,762	0,975
client_group	5 vs Others	*2,379	1,331	1,104	*1,384	*1,512	*1,826	1,407
client_group	6 vs Others	*1,760	*1,570	0,964	1,129	*1,454	1,032	*1,590
client_group	7 vs Others	*0,237	*0,655	0,979	*0,554	*0,658	*0,243	*0,390
client_group	8 vs Others	0,843	0,911	1,256	1,188	0,949	0,696	1,220
client_group	9 vs Others	*1,934	*1,966	1,283	1,025	1,140	0,870	*1,626
client_group	10 vs Others	*1,704	*1,878	1,275	*1,395	*1,652	0,945	1,252
client_group	11 vs Others	*2,982	*3,016	*3,129	1,225	*1,916	1,349	1,340
client_group	12 vs Others	*1,633	1,510	1,158	1,001	*1,382	1,128	*1,564
client_group	13 vs Others	*1,689	0,843	0,668	0,964	*1,298	0,921	1,637
client_group	14 vs Others	*1,664	*1,641	*1,677	1,210	*1,436	1,015	*1,758
client_group	15 vs Others	*1,695	*1,738	1,359	1,026	*1,383	0,971	*1,596
client_group	16 vs Others	0,935	0,908	1,152	0,985	*1,239	0,950	0,924
client_group	17 vs Others	*0,434	0,776	1,274	*0,405	*0,657	*0,165	*0,627
client_group	18 vs Others	*2,575	*2,331	*1,867	1,243	*2,014	*1,753	*1,488
client_group	19 vs Others	*1,660	*1,470	*1,564	1,095	*1,630	1,048	*1,610
client_group	20 vs Others	*2,218	*1,491	*1,450	*1,943	*1,800	*1,743	*1,740
client_group	21 vs Others	*2,611	1,633	1,232	*2,207	*2,290	0,988	1,090
client_group	22 vs Others	1,300	1,044	0,810	0,918	1,183	*1,737	*1,750
client_group	23 vs Others	*1,639	*1,598	1,335	1,015	*1,462	1,021	*1,512
client_group	24 vs Others	1,157	1,018	1,106	1,033	*1,231	0,789	1,150

Les rapports de cotes marqués d'un astérisque (*) sont significatifs au seuil de signification de 5%.

avons pris soin de modifier l'échelle des variables afin que les rapports de cotes aient le plus de sens possible. Ainsi, la prime et le revenu ont étés convertis en milliers de dollars ; l'effet de la variable est donc ramené pour chaque variation de 1000$. Il ne fait aucun doute que, pour une variation de prime de 1$, le rapport de cotes lié à ces variables aurait été environ égal à 1,000. C'est pourquoi nous avons pris soin d'ajouter la colonne « unités » au tableau 5.7, qui représente la variation par rapport à laquelle le rapport de cotes doit être interprété. Il faut donc faire attention à l'interprétation des rapports de cotes. À titre d'exemple, l'effet de la variable client_age_owner1 peut sembler faible ; mais il s'agit là de l'impact de la variable pour une variation d'une seule année. Le rapport de cotes de cette variable aurait été plus impressionnant si la variable représentait une variation de 10 années.

Par ailleurs, nous émettons également une mise en garde en ce qui concerne la significativité des rapports de cotes. Bien que certains ne soient pas statistiquement significatifs, il ne faut pas penser que les variables sous-jacentes soient totalement dépourvues d'intérêt et qu'elles ne contribuent pas à la mesure d'espérance de vie. En effet, bien que les résultats intérimaires puissent apporter des explications intéressantes sur le comportement des clients, le but ultime de la démarche est de calculer une espérance de vie. Dans ce contexte, la significativité ou la non significativité d'une variable n'est pas ce qui nous importe le plus ; que les rapports de cotes liés à une variable ne soient pas tous significatifs n'empêche pas que cette variable puisse apporter une contribution, même mineure, à l'espérance de vie. Par ailleurs, le seul fait que les variables aient été retenues dans la sélection de variable de type pas-à-pas est une justification suffisante pour les conserver toutes.

5.3.2. Utilisation de sous-ensembles aléatoires des échantillons d'entraînement et de validation

Les calculs matriciels des sous-sections 5.3.3 et 5.3.5 ont été réalisés avec le logiciel R version 2,10. Les capacités de R en termes d'importation de données étant assez limitées et le nombre de matrices à importer étant assez élevé (le nombre de clients à prédire multiplié par le nombre d'années de prédiction), des problèmes de mémoire survenaient. Après vérifications, il nous a semblé que le problème était relié aux fonctions d'importations qui sont disponibles dans les bibliothèques de fonctions par défaut de R (problème d'allocation de mémoire pour des ensembles de données volumineux). Il existe des bibliothèques permettant d'effectuer ces manipulations sans difficulté. Toutefois, elles n'étaient pas disponibles dans la distribution R installée sur les postes de travail de la compagnie dans laquelle le projet de ce mémoire a été réalisé. Nous ne sommes pas parvenus à obtenir ces bibliothèques dans les délais requis pour ce mémoire. Nous nous sommes donc limités à des sous-ensembles de 12 000 clients sélectionnés aléatoirement dans chacun des ensembles d'entraînement et de validation. Nous avions donc 12 000 clients sur lesquels obtenir la courbe de survie et l'espérance de vie dans l'échantillon d'entraînement ainsi que 12 000 autres pour ces mêmes quantités dans l'échantillon de validation. Bien que ces effectifs soient relativement petits par rapport à ceux disponibles dans l'échantillon d'entraînement (151 414 clients) et de validation (64 891 clients), le fait que 12 000 soit un nombre, en absolu, très grand et qu'ils aient étés choisis aléatoirement devrait nous garantir que les courbes de survies et les espérances de vie obtenues soient très près de la valeur sur l'ensemble des échantillons respectifs. Pour une utilisation réelle de la méthode, il est évident que nous suggérons d'employer l'ensemble des individus pour lesquels nous désirons obtenir une espérance de vie.

5.3.3. Matrices de transition à k pas

Une fois que les régressions logistiques multinomiales ont été ajustées, il nous faut obtenir les espérances de vie. Il faut donc dans un premier temps obtenir les estimateurs des matrices de transition à k pas pour ensuite obtenir la courbe de survie prédite, de laquelle nous pouvons obtenir l'espérance de vie. Dans cette section, nous utilisons la méthode proposée par McFarland (1970) et Spilerman (1972) afin d'obtenir l'estimateur de la matrice de transition à k pas donné par l'équation (2.5.2). Rappelons que cette méthode suppose que chaque individu est une chaîne de Markov. Bien que le processus au niveau de la population ne soit pas nécessairement markovien, il admet une distribution qui est donnée dans l'équation (2.5.2). Nous avons présenté dans le tableau 5.9 les valeurs des premières matrices de transition ainsi obtenues. Bien que nous ayons effectué les prédictions pour un total de 70 années, nous avons limité la présentation des matrices de transition à ces cinq premières années par souci de concision. De plus, afin d'alléger les tableaux, nous nous sommes limités aux matrices obtenues sur l'échantillon de validation, auxquelles nous avons juxtaposé les matrices obtenues par la méthode de Markov simple. Les différences constatées entre ces paires de matrices sont attribuables aux caractéristiques des modèles utilisés, soit le modèle de Markov simple (voir section 2.2) et le modèle de McFarland-Spilerman (voir section 2.5).

5.3.4. Sélection de la fonction d'ajustement pour le modèle semi-paramétrique

Lors de l'ajustement des régressions logistiques multinomiales, nous avons constaté que l'effet de la variable client_since qui découle d'une modélisation sur une année et utilisant l'ensemble des clients ne correspondait pas à l'effet réel de cette variable sur le temps de vie d'un client ; c'est pour cette raison que nous avons décidé d'utiliser une fonction d'ajustement (voir section 4.2).

TAB. 5.9: Matrices de transitions à k de la méthode de McFarland-Spilerman et de la méthode de Markov simple des 5 premières années de prédiction (obtenues sur l'échantillon de validation)

Temps (k, en années)	Méthode McFarland-Spilerman	Méthode Markov simple
1	$\begin{pmatrix} 0{,}8907 & 0{,}0270 & 0{,}0221 & 0{,}0592 \\ 0{,}1226 & 0{,}7390 & 0{,}0000 & 0{,}1384 \\ 0{,}2009 & 0{,}0000 & 0{,}7597 & 0{,}0394 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$	$\begin{pmatrix} 0{,}9164 & 0{,}0200 & 0{,}0171 & 0{,}0464 \\ 0{,}1223 & 0{,}7098 & 0{,}0000 & 0{,}1678 \\ 0{,}0937 & 0{,}0000 & 0{,}7682 & 0{,}1381 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$
2	$\begin{pmatrix} 0{,}8124 & 0{,}0396 & 0{,}0363 & 0{,}1107 \\ 0{,}1951 & 0{,}5627 & 0{,}0027 & 0{,}2395 \\ 0{,}3091 & 0{,}0049 & 0{,}6072 & 0{,}0788 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$	$\begin{pmatrix} 0{,}8439 & 0{,}0326 & 0{,}0288 & 0{,}0947 \\ 0{,}1989 & 0{,}5063 & 0{,}0021 & 0{,}2927 \\ 0{,}1578 & 0{,}0019 & 0{,}5918 & 0{,}2485 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$
3	$\begin{pmatrix} 0{,}7523 & 0{,}0457 & 0{,}0450 & 0{,}1561 \\ 0{,}2405 & 0{,}4371 & 0{,}0060 & 0{,}3163 \\ 0{,}3783 & 0{,}0103 & 0{,}4949 & 0{,}1165 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$	$\begin{pmatrix} 0{,}7801 & 0{,}0401 & 0{,}0365 & 0{,}1433 \\ 0{,}2445 & 0{,}3634 & 0{,}0050 & 0{,}3872 \\ 0{,}2003 & 0{,}0045 & 0{,}4573 & 0{,}3379 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$
4	$\begin{pmatrix} 0{,}7044 & 0{,}0483 & 0{,}0501 & 0{,}1963 \\ 0{,}2689 & 0{,}3457 & 0{,}0092 & 0{,}3762 \\ 0{,}4222 & 0{,}0153 & 0{,}4103 & 0{,}1523 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$	$\begin{pmatrix} 0{,}7232 & 0{,}0441 & 0{,}0414 & 0{,}1913 \\ 0{,}2690 & 0{,}2628 & 0{,}0080 & 0{,}4602 \\ 0{,}2269 & 0{,}0072 & 0{,}3548 & 0{,}4111 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$
5	$\begin{pmatrix} 0{,}6649 & 0{,}0489 & 0{,}0531 & 0{,}2322 \\ 0{,}2862 & 0{,}2777 & 0{,}0120 & 0{,}4240 \\ 0{,}4493 & 0{,}0194 & 0{,}3455 & 0{,}1859 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$	$\begin{pmatrix} 0{,}6721 & 0{,}0458 & 0{,}0441 & 0{,}2380 \\ 0{,}2794 & 0{,}1920 & 0{,}0108 & 0{,}5179 \\ 0{,}2421 & 0{,}0097 & 0{,}2764 & 0{,}4719 \\ 0{,}0000 & 0{,}0000 & 0{,}0000 & 1{,}0000 \end{pmatrix}$

Les densités qui ont été considérées pour obtenir notre fonction d'ajustement sont la loglogistique, la lognormale et la Weibull (voir annexe A). Les paramètres de ces densités ont été estimés par la méthode du maximum de vraisemblance à l'aide du logiciel Matlab version 8 ; les données utilisées à cette fin sont celles de la distribution cumulative de la variable `client_since`. Les estimateurs des paramètres obtenus et la logvraisemblance des modèles ajustés sont présentés dans le tableau 5.10. Nous avons comparé la qualité de l'ajustement en comparant la logvraisemblance de chacun des trois modèles ajustés, lesdits modèles étant simplement l'ajustement des fonctions de survie découlant de chacune des densités ci-mentionnées sur la distribution cumulative de la variable client_since. Nous avons obtenu que le meilleur ajustement était

TAB. 5.10: Ajustement des courbes sur la distribution de la variable client_since

Famille de densité	Paramètres estimés	Logvraisemblance
Lognormale	$\mu = 4{,}1745$, $\sigma = 1{,}5553$	$-71\,664$
Weibull	$a = 68{,}785$, $b = 1{,}136$	$-73\,141$
Loglogistique	$\mu = 4{,}0218$, $\sigma = 0{,}8220$	$-72\,825$

celui de la lognormale (voir tableau 5.10). La représentation graphique des différentes courbes de survie ajustées versus la courbe de survie empirique (de la variable client_since) se retrouve à la figure 5.4. À la vue de ce graphique, la qualité de l'ajustement peut sembler discutable. Nous attribuons cette situation à la présence d'un taux de censure très élevé ; en effet, il est de 90,2% à la fin de la première année. Il s'agit ici d'une contrainte inhérente aux données et au problème ; l'ajustement du modèle sur des clients d'un assureur présentant une moins bonne rétention (ou, alternativement, en limitant la modélisation aux nouveaux clients) présenterait un taux de censure substantiellement moins élevé . Nous conservons nénamoins cette approche puisque nous désirions une méthode objective permettant de tenir compte d'un effet plus réaliste de la variable client_since.

5.3.5. Courbe de survie prédite et espérance de vie

Maintenant que nous disposons des matrices de transition à k pas et de la fonction d'ajustement, nous pouvons obtenir la courbe de survie prédite à l'aide du modèle semi-paramétrique dérivé de la méthode de McFarland-Spilerman. À la figure 5.5, nous retrouvons la courbe représentant la fonction d'ajustement, la courbe de survie obtenue à partir des matrices de transitions à k pas et la courbe de survie ajustée, qui est simplement le produit des deux courbes précédentes. Afin de visualiser l'espérance de vie, nous avons coloré la zone la représentant sur les graphiques. Bien que les courbes puissent sembler lisses, il importe de se rappeler qu'il ne s'agit que de points reliés, notre modèle utilisant des valeurs discrétisées du temps. Afin de faciliter la comparaison des

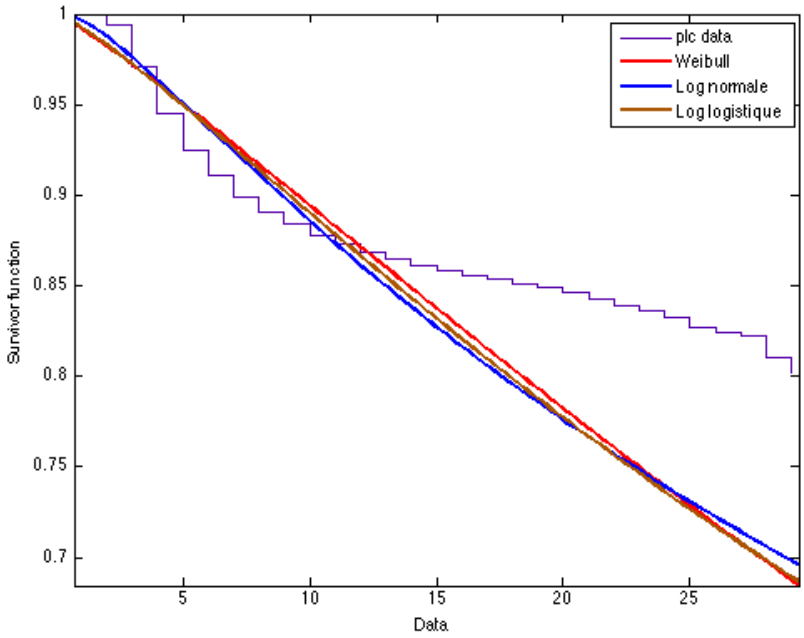
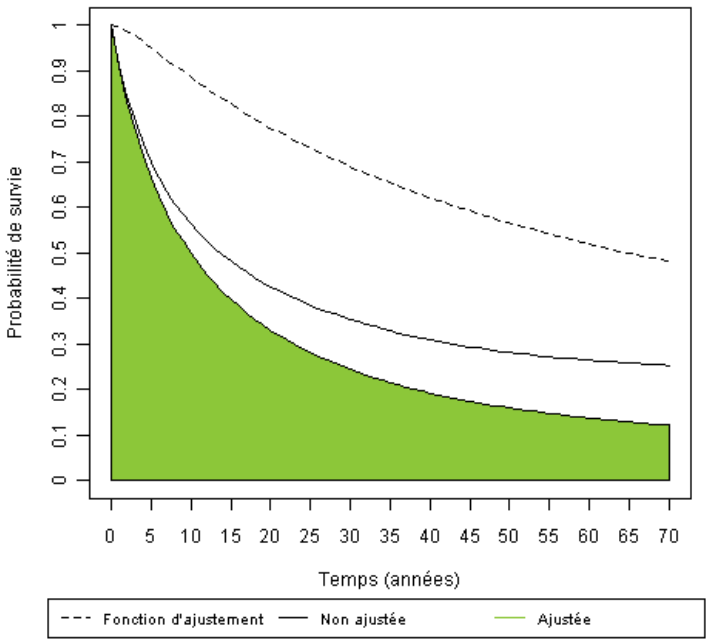


FIG. 5.4: Ajustement des courbes de survie pour l'obtention de la fonction d'ajustement

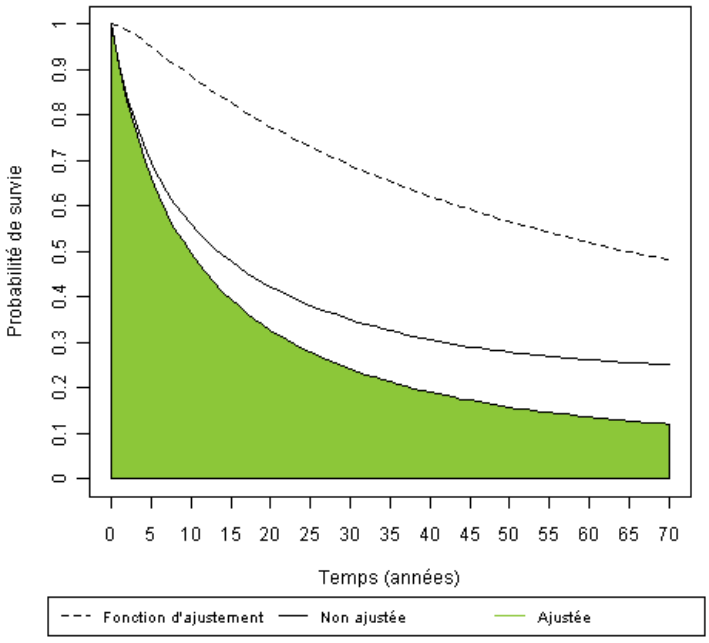
TAB. 5.11: Espérances de vie du modèle de McFarland-Spilerman avec ajustement

Échantillon	n	Espérance de vie (années)
Entraînement	12 000	23,36
Validation	12 000	23,16

modèles, nous avons représenté la courbe de survie prédite du modèle de Markov simple et la courbe de survie dérivée de McFarland-Spilerman avec ajustement de l'échantillon de validation à la figure 5.6. Notons que les aires sous la courbe des graphiques de la figure 5.5 valent respectivement 23,36 et 23,16, qui correspondent aux espérances de vie ainsi prédites (voir tableau 5.11).



(a) Entraînement



(b) Validation

FIG. 5.5: Courbes de survies prédites par la méthode de McFarland-Spilerman

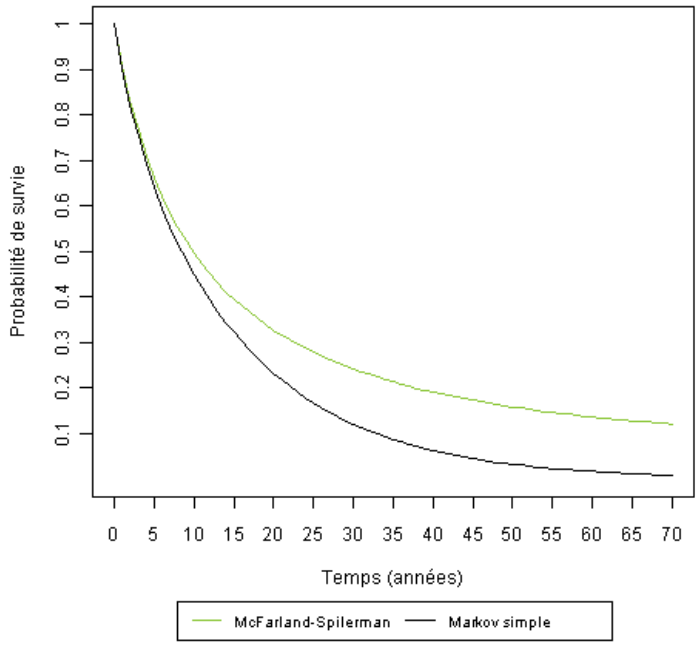


FIG. 5.6: Courbes de survie prédites du modèle de Markov simple et du modèle de McFarland-Spilerman avec ajustement (échantillon de validation)

5.3.5.1. *Censure au-delà de 70 ans au sein de la compagnie*

En regardant de près les graphiques de la figure 5.5, nous constatons que les courbes cessent après 70 ans. Il ne s'agit pas de la limite de représentation des courbes, mais bien de la limite imposée au modèle. En effet, bien que l'utilisation de la fonction d'ajustement permette de corriger en partie la tendance à long terme en bornant la probabilité de survie, il semble que la tendance voulant que la probabilité de survie augmente avec le temps soit assez forte pour qu'après 70 ans, la probabilité de survie globale soit d'un peu plus de 10%. Il aurait donc été possible d'effectuer les prédictions pour les temps ultérieurs. Nous avons décidé qu'à partir de ce point, nous considérions les individus restant comme censurés et que nous calculions l'espérance de vie en tenant compte de la troncature à droite impliquée (voir chapitre 4). Il aurait été possible de cesser les prédictions après un nombre d'années inférieur ou

supérieur ; mais la différence sur l'estimateur aurait été très faible étant donné que le calcul de l'espérance de vie tient compte de la troncature des unités restantes.

5.4. Approche par simulations

Le dernier modèle qui a été implémenté pour l'obtention de l'espérance de vie est un modèle par simulation. Nous avons obtenu les espérances de vie de ce modèle à la fois pour les jeux de données d'entraînement et de validation complets et pour les mêmes sous-ensembles de 12 000 clients que ceux utilisés pour le modèle de McFarland-Spilerman avec ajustement. Le tableau 5.12 présente ces espérances de vie de même que l'écart-type de simulation. Notons que l'écart-type de simulation est l'écart-type qui découle du fait d'avoir simulé plusieurs fois les chemins empruntés par les clients à travers les états. Un grand écart-type implique donc une grande variabilité du résultat d'espérance de vie obtenu lors des simulations et inversement.

La première chose que nous constatons en regardant le tableau 5.12 est que les valeurs d'espérance de vie obtenues sur les sous-ensembles de 12 000 individus sont près des valeurs obtenues avec le modèle de McFarland-Spilerman ajusté. En effet, pour le même sous-ensemble de 12 000 individus, l'espérance de vie par la méthode de McFarland-Spilerman était de 23,36 (23,16 pour la validation) alors qu'elle est de 23,14 (22,99 pour la validation) dans l'approche par simulation. Les écarts-types de simulation ne semblent pas indiquer que ces disparités soient surprenantes puisqu'il y a moins de deux écarts-types de simulation entre les valeurs obtenues par simulation et les valeurs obtenues par la méthode de McFarland-Spilerman. En effet, les écarts-types de simulation sont respectivement de 0,18 pour l'échantillon d'entraînement et de 0,17 pour l'échantillon de validation. En ce sens, rien ne semble indiquer que les deux méthodes soient statistiquement différentes.

La deuxième chose que nous constatons du tableau 5.12 est que les espérances de vie varient légèrement selon qu'elles sont calculées sur l'ensemble des individus des échantillons ou sur des sous-ensembles d'individus choisis aléatoirement. En faisant fi de l'erreur de simulation, cette disparité est tributaire de la sélection aléatoire des clients, les 12 000 clients sélectionnés ne pouvant, à eux seuls, représenter parfaitement l'ensemble de leurs semblables. Il s'agit donc d'une erreur d'échantillonnage. Dans ce mémoire, nous n'avons pas cherché à quantifier cette erreur d'échantillonnage. Il aurait néanmoins été possible d'annuler cette erreur d'échantillonnage en effectuant une sélection aléatoire des clients (sous-ensembles de 12 000 individus) un nombre élevé de fois et de prendre la moyenne des espérances de vie obtenues à chacun des calculs. Bien que le temps de simulation ne soit pas faramineux (en-deçà d'une demi-heure pour les calculs par simulation avec 100 simulations par individus), sa répétition un nombre suffisant de fois pour annuler l'erreur d'échantillonnage rendrait cette approche inutile en pratique. Il importe de garder en mémoire l'objectif premier de la méthode développée dans ce mémoire, soit la prédiction de sous-ensembles d'individus : tant que ces sous-ensembles ne dépassent pas 12 000 individus, l'implémentation de la méthode de McFarland-Spilerman est exacte. Dans le cas où la taille de ces sous-ensembles est supérieure à 12 000, l'utilisation de la méthode exacte de McFarland-Spilerman implique d'accepter cette erreur d'échantillonnage, l'alternative étant de troquer cette erreur d'échantillonnage par l'erreur de simulation de l'approche par simulation.

La dernière chose que nous constatons du tableau 5.12 est que les écarts-types de simulation sont substantiellement plus faibles lorsque les tailles échantillonnales sont plus élevées. La raison est tout simplement que l'espérance de vie d'un échantillon est calculée comme une moyenne de l'espérance de vie

TAB. 5.12: Espérances de vie du modèle par simulation (N = 100 simulations)

Échantillon	n	Espérance de vie (années)	Écart-type de simulation
Entraînement	12 000	23,14	0,18
Validation	12 000	22,99	0,17
Entraînement	151 414	23,01	0,06
Validation	64 891	23,08	0,08

de chacun des clients constituant cet échantillon. Ainsi, plus le nombre d'individus est grand, plus stable est la mesure entre les simulations, les écarts de comportements individuels ayant une incidence moindre sur la mesure globale.

5.5. RÉSULTATS DÉRIVÉS DU MODÈLE

Il est possible, grâce à notre modèle, d'obtenir des espérances de vie pour certains sous-ensembles de la population. Ces sous-ensembles peuvent être définis par des variables qui composent notre modèle (tel que le groupe du client, son âge, son sexe, etc.) ou encore provenir de sources externes et présenter une grande variabilité en termes des variables présentes. Notre mise en garde concerne le premier cas. En effet, lorsque de telles comparaisons sont effectuées (par exemple : les hommes âgés de plus de 65 ans versus les femmes âgées de moins de 65 ans faisant partie du groupe d'affinité portant le numéro 9), il importe de conserver en mémoire que l'espérance de vie prédite se fonde sur le comportement des clients à l'intérieur d'une seule année. Bien que le modèle puisse exploiter l'information de l'ensemble des clients afin de modéliser l'effet de variables sur l'évolution du temps de vie (tel que l'âge du client) et ainsi prédire les tendances futures, il ne peut se prémunir de comportements de clients qui sont causés par des événements ponctuels non expliqués par les variables du modèle. Le cas le plus fréquent en assurance est celui où des modifications aux primes sont apportées à certains groupes de clients. Dans ce contexte, nous nous attendons à ce que ces clients quittent la compagnie à un

taux plus élevé que leurs semblables. Inversement, il est possible que certains sous-ensembles aient une rétention plus élevée pendant l'année d'étude parce qu'ils ont été particulièrement ciblés par des efforts de marketing. En conclusion, il est tout-à-fait possible de comparer les espérances de vie des clients de différents groupes ; mais cela doit se faire en conservant en mémoire que l'hypothèse fondamentale est alors que le comportement des clients n'ait pas été altéré pendant l'année de modélisation par des facteurs exogènes non expliqués par les variables et propres à ces sous-ensembles de clients.

Il serait possible d'inclure des variables indicatrices dans les modélisations afin de tenir compte de ces différents événements. Cela pose toutefois un certain nombre de contraintes pratiques. Premièrement, cela nécessite une connaissance approfondie du passé de l'entreprise afin de reconnaître et d'inclure les événements pertinents. Deuxièmement, cela exige des efforts de programmation susbtantiels afin d'identifier les clients concernés par ces événements. Troisièmement, bien que le nombre d'individus disponibles pour la modélisation soit élevé, il importe de se rappeler que plus il y a de variables catégorielles dans les régressions logistiques multinomiales, plus grand est le risque que les modèles ne convergent pas ou que les effets détectés soient discutables. Quatrièmement, une dernière contrainte pratique apparaîtrait lorsque nous tenterions d'effectuer des prédictions : il faudrait alors fournir des valeurs futures des variables que nous avons incluses dans le modèle afin d'être en mesure d'obtenir une espérance de vie. Il faudrait ainsi anticiper certaines modifications aux primes ou encore planifier à l'avance quels clients seront ciblés par certaines campagnes marketing. En d'autres termes, il faudrait alors incorporer un nombre important de considérations reliées à l'entreprise dans l'obtention de la mesure d'espérance de vie, ce qui complexifie beaucoup son obtention. Pour toutes ces raisons, il a été décidé de ne pas introduire de telles

variables dans nos modélisations. À titre d'exemple, il n'aurait pas été particulièrement complexe d'inclure une variable représentant la variation de la prime par rapport à l'année précédente (à la différence de l'identification des clients ciblés par certaines campagnes marketing ou de modifications ciblant certains groupes d'affinité seulement). Il a cependant été jugé non souhaitable que la mesure d'espérance de vie prédite dépende de considérations non statistiques.

5.6. Validation des mesures

Maintenant que nous avons obtenu des espérances de vie grâce à nos modèles, il nous faut obtenir des mesures de validation de ces quantités. Tel que discuté à la section 4.3, la nature même du problème fait en sorte que nous ne pourrons jamais savoir si les quantités prédites sont exactes étant donné qu'il nous est impossible d'attendre que l'ensemble des clients soient décédés afin de valider la mesure. Néanmoins, il est possible d'obtenir des quantités nous donnant des mesures de performances de nos modèles. La première de ces mesures est l'espérance de vie « censurée après cinq ans ». Nous définissons l'espérance de vie censurée comme l'espérance de vie calculée en assumant que les individus vivent au plus cinq années ; cela est équivalent à tronquer la courbe de survie au-delà de cinq ans. En d'autres termes, cela est équivalent à cesser les prédictions après cinq années et de ne *pas* corriger pour la censure à droite. Tel que mentionné à la fin de la section 5.1, nous avons conservé quatre années additionnelles indépendantes de l'année sur laquelle a été effectuée la modélisation ; en effectuant nos prédictions à partir du temps initial, nous disposons d'un total de cinq années pour nos validations. Les valeurs d'espérance de vie censurée après cinq ans sont présentées dans la tableau 5.13. La première chose que nous constatons de ce tableau est que le modèle de Markov simple semble bien se comporter à court terme. En effet, l'espérance de vie prédite est de 3,99 années alors que la valeur empirique est de 4,02. Comment expliquer

un si faible écart entre les valeurs prédites par le modèle de Markov simple et les deux autres approches sur une période de cinq années alors que la valeur prédite sur l'ensemble de la vie des clients était presque deux fois moindre (13,38 ans plutôt qu'environ 23 années pour les autres méthodes) ? Nous pensons que pour une courte période, le modèle de Markov simple puisse être une bonne approximation d'une chaîne de Markov dans laquelle les probabilités de transition sont non stationnaires ou hétérogènes, mais que ces hypothèses deviennent moins réalistes pour une période plus longue. La deuxième chose que nous constatons à partir du tableau 5.13 est que l'ajustement des modèles à l'aide de la fonction d'ajustement (McFarland-Spilerman, approche par simulation calculée sur les sous-ensembles de 12 000 clients et approche par simulation calculée sur l'ensemble des échantillons disponibles) semble pertinent, l'ajustement permettant de rapprocher les espérances de vie prédites non ajustées de la valeur empirique sans non plus sur-corriger.

Afin d'évaluer la performance de notre mesure d'espérance de vie, nous avons utilisé une deuxième quantité. Il s'agit de la probabilité de survie après cinq ans. Le tableau 5.14 présente ces probabilités de survie empiriques et celles obtenues par la méthode de McFarland-Spilerman. Le tableau 5.15 présente ces mêmes quantités calculées chez les individus qui sont en réalité (cas empirique) en vie après cinq années et chez ceux qui sont en réalité décédés. Du premier tableau nous constatons que nos prédictions semblent raisonnables, les valeurs prédites par la méthode de McFarland-Spilerman étant près des valeurs empiriques. Ceci n'est que le reflet de la conclusion obtenue avec l'espérance de vie censurée après cinq années. C'est du second tableau que nous effectuons les observations les plus intéressantes. En effet, l'objectif est de prédire des probabilités de survie élevées chez les individus qui sont en réalité

TAB. 5.13: Espérance de vie censurée après cinq ans

Espérance de vie censurée	n		Ensemble de données	
	Entraînement	Validation	Entraînement	Validation
Markov simple	216 305		3,99	
McFarland-Spilerman sans ajustement	12 000	12 000	4,14	4,13
McFarland-Spilerman avec ajustement	12 000	12 000	4,06	4,05
Approche par simulation sans ajustement (N = 100 sim.)	12 000	12 000	4,06	4,05
Approche par simulation avec ajustement (N = 100 sim.)	12 000	12 000	3,98	3,98
Approche par simulation sans ajustement (N = 100 sim.)	151 414	64 891	4,06	4,06
Approche par simulation avec ajustement (N = 100 sim.)	151 414	64 891	3,98	3,98
Empirique	12 000	12 000	4,02	4,01
Empirique	151 414	64 891	4,01	4,02

en vie après cinq années et faibles chez les individus qui sont en réalité décédés après ces cinq années. Si notre méthode était parfaite (ce qui est utopique d'un point de vue statistique), nous aurions des probabilités de survie prédites de 100% chez les individus qui sont en vie après cinq ans et des probabilités de survie nulles chez les individus qui sont en réalité décédés après cinq ans. Pour la méthode de McFarland-Spilerman, nous avons obtenu une probabilité prédite de survie de 0,7155 sur l'échantillon d'entraînement et de 0,7118 sur l'échantillon de validation pour les individus qui sont réellement en vie et de 0,5606 (0,5686 sur l'échantillon de validation) pour les individus qui sont en réalité décédés après cinq ans. Il est intéressant de contraster ces valeurs avec les probabilités de survie prédites globales, qui sont de 0,6808 sur l'échantillon d'entraînement et de 0,6733 sur l'échantillon de validation. Notre méthode permet d'effectuer une discrimination certaine et dans le bon sens des individus en fonction de leur temps de vie. Il pourrait sembler que cette différence en termes de discrimination soit faible. Toutefois, il importe de se rappeler deux choses. Premièrement, cette différence est calculée sur les cinq premières années seulement et devrait s'accroître pour des prédictions plus grandes, les clients fidèles ayant tendance à demeurer au même état. Deuxièmement, l'objectif de la méthode proposée est d'obtenir une bonne espérance de vie *en moyenne* pour un sous-ensemble d'individus ; il est possible d'obtenir une bonne espérance de vie du point de vue d'un échantillon sans que ces prédictions ne soient parfaites du point de vue de chacun des individus de cet échantillon.

Notons qu'il est théoriquement possible d'obtenir ces quantités pour l'approche par simulation. Toutefois, pour des raisons techniques, nous ne les

TAB. 5.14: Probabilités de survie après cinq ans

Échantillon	Statut prédit après 5 ans	Empirique		McFarland-Spilerman	
		N	Probabilité	N	Probabilité
Entraînement	En vie	8 169	0,6808	7 992	0,6660
n = 12 000	Décédés	3 831	0,3192	4 008	0,3340
Validation	En vie	8 084	0,6737	7 980	0,6650
n = 12 000	Décédés	3 916	0,3263	4 020	0,3350

TAB. 5.15: Probabilités de survie après cinq ans en fonction du statut réel après cinq ans

Échantillon	Statut réel après 5 ans	McFarland-Spilerman	Empirique
Entraînement	En vie	0,7155	-
	Décédés	0,5606	-
n = 12 000	Total	0,6660	0,6808
Validation	En vie	0,7118	-
	Décédés	0,5686	-
n = 12 000	Total	0,6650	0,6733

avons pas calculées. Par ailleurs, la similitude que nous avons constatée précédemment entre les résultats de l'approche par simulation et ceux de la méthode de McFarland-Spilerman nous garantit des résultats similaires pour ces mesures.

Dans ce chapitre, nous avons présenté les données disponibles pour notre modélisation du temps de vie pour ensuite présenter les résultats des trois approches utilisées. Nous avons constaté que le modèle de Markov simple a l'avantage d'être simple et qu'il semble donner de bons résultats à court terme. Cependant, le fait qu'il ne puisse tenir compte de la décroissance des probabilités de décès dans le temps entraîne des disparités avec les valeurs obtenues par les deux autres approches. Le modèle basé sur la méthode de McFarland-Spilerman avec ajustement permet d'obtenir efficacement une valeur d'espérance de vie à l'aide d'une formule explicite. Toutefois, la transition d'un logiciel statistique à un autre entraîne certains désagréments et nous avons eu des

difficultés à importer nos jeux de données volumineux avec le logiciel R. L'approche par simulation remédie à ce problème et permet d'obtenir une mesure d'espérance de vie basée sur des scénarios aléatoires répétés ; le temps de calcul s'accroît alors. Ainsi, les trois méthodes ont des avantages et pourraient être envisagées. L'absence de régressions du modèle de Markov simple le rend très facile à implémenter et diminue substantiellement les manipulations requises sur les jeux de données. Cette méthode pourrait donc être envisagée pour des valeurs comparatives d'espérances de vie de différents sous-ensembles de clients. Les deux autres approches, quoique plus complexes, nous semblent de loin préférables étant donné qu'elles permettent de tenir compte de la diminution des probabilités de décès au fil du temps en exploitant au maximum l'hétérogénéité présente dans les jeux de données disponibles. De ces deux dernières approches, laquelle préférer ? Il nous semble que ce choix dépend de contraintes plus pratiques que théoriques, les deux méthodes donnant des résultats similaires.

CONCLUSION

Dans ce mémoire, nous avons proposé et comparé trois approches permettant d'effectuer des prédictions de l'espérance de vie de clients pouvant posséder différents types de produits d'assurance. Ces trois approches sont des modèles par chaînes de Markov ; l'un effectue l'hypothèse d'homogénéité des probabilités de transition, les deux autres en supposent l'hétérogénéité. Dans ces deux derniers cas, les probabilités de transition ont été estimées par des régressions logistiques multinomiales. De nos calculs, nous avons constaté que les modèles par chaînes de Markov constituent une classe de modèles flexibles et qui nous permettent de tenir compte des multiples situations possibles des clients d'une façon simple et intuitive.

En ce qui concerne les résultats obtenus, les espérances de vie prédites oscillent entre 22,99 années et 23,36 années pour les méthodes de McFarland-Spilerman et l'approche par simulation, les différences étant attribuables aux sous-ensembles de clients considérés et aux erreurs de simulation. Le modèle de Markov simple nous a quant à lui donné une espérance de vie moyenne de 13,38 années. Nous pensons que les disparités entre cette méthode et les deux précédentes sont dues au fait que ses hypothèses, trop fortes, ne sont pas respectées par notre problème. Notons qu'à courts termes, ce modèle semble toutefois être une bonne approximation de notre situation, les valeurs obtenues pour une projection sur une période de cinq années étant très près des valeurs empiriques.

De multiples voies d'amélioration sont envisageables. En ce qui concerne les modèles markoviens, il serait intéressant d'étudier les modèles par chaînes de Markov à temps continus, ce qui éviterait de devoir discrétiser la variable temps et permettrait de bénéficier pleinement de l'information disponible. Du côté de la modélisation des probabilités de transition, il serait intéressant d'essayer d'utiliser des modèles plus avancés que la régression logistique multinomiale tel que les modèles de valeurs extrêmes généralisées ou encore des modèles de régression logistique multinomiales mixtes. Du côté de l'espérance de vie, nous notons que les méthodes de ce mémoire permettent d'obtenir une espérance de vie tronquée à gauche au temps présent (temps auquel les modèles sont ajustés) et tenant compte de la troncature à droite dans les prédictions. Le fait que nous ayons corrigé pour la troncature à droite implique que l'espérance de vie obtenue donne une mesure du temps moyen jusqu'à ce que tous les clients quittent la compagnie : ceci se fait simplement en divisant l'espérance de vie calculée par la proportion d'individus qui ne sont pas décédés lorsque nous décidons d'arrêter les calculs. La troncature à gauche, pour laquelle aucune correction n'est apportée, implique que notre estimateur de l'espérance de vie doit être interprété comme le temps de séjour moyen qu'il reste aux clients *à partir du moment où il est calculé*. En d'autres termes, nous calculons l'espérance de vie conditonnelle au fait que les clients sont en vie au temps présent et sans égards au temps déjà passé au sein de la compagnie (à l'exception près de la variable explicative *client_since*). Il serait possible d'obtenir un estimateur de l'espérance de vie des clients représentant le temps de vie total des clients, c'est-à-dire tenant compte de la troncature à gauche. Notons que la raison pour laquelle nous n'avons pas corrigé pour la troncature à gauche est simplement que c'est la quantité que nous désirions obtenir ; le fait d'en

tenir compte pourrait toutefois apporter une information additionnelle pertinente pour la compagnie. Du côté plus pratique, il serait intéressant d'ajuster les modèles présentés dans ce mémoire sur le marché direct, qui constitue une clientèle beaucoup plus friable. Cette quantité accrue de variabilité devrait permettre une meilleure discrimination entre les individus fidèles et les individus moins fidèles, le taux de censure y étant substantiellement moins élevé. Enfin, dans le futur, la compagnie pourrait vouloir intégrer à cette mesure d'autres types d'assurance, dont l'assurance-vie et l'assurance-voyages, le modèle développé dans ce mémoire étant suffisamment flexible pour pouvoir tenir compte de tels ajouts : il suffirait simplement d'ajouter des états à notre modèle markovien.

Bibliographie

Anderson, T. W. et Goodman, L. A. (1957). Statistical inference about markov chains. *The Annals of Mathematical Statistics*, **28**(1), 89–110.

Chipman, J. S. (1960). The foundations of utility. *Econometrica*, **28**(2), 193–224.

Cohen, H. (2011). *Numerical Approximation Methods*. Springer.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2), 187–220.

Czado, C. et Rudolph, F. (2002). Application of survival analysis methods to long-term care insurance. *Insurance : Mathematics and Economics*, **31**(3), 395–413.

Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, **56**(296), 841–868.

Hesselager, O. (1994). A markov model for loss reserving. *Astin Bulletin*, **24**(2), 183–194.

Hodge, R. W. (1966). Occupational mobility as a probability process. *Demography*, **3**(1), 19–34.

Klein, J. P. et Moeschberger, M. L. (1997). *Survival analysis : techniques for Censored and Truncated Data*. Springer.

Kutner, M., Nachtsheim, C., Neter, J., et Li, W. (2004). *Applied linear statistical models, 5th edition*. McGraw-Hill.

Kwon, H. S. et Jones, B. L. (2008). Applications of a multi-state risk factor/mortality model in life insurance. *Insurance : Mathematics and Economics*,

43(3), 394–402.

Luce, R. D. (1959). *Individual Choice Behavior : A Theoretical Analysis*. Wiley.

Makhzoum, S. (2002). *Analysis of policy cancellations at Meloche Monnex*. Université de Montréal.

McFadden, D. (1973). *Conditional Logit Analysis of Qualitative Choice Behavior dans P. Zarembka,* Frontiers in Econometrics, chap. 4. New York : Academic Press, pp. 105–142.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, **3**, 303–328.

McFadden, D. (1980). Econometric models for probabilistic choice among products. *The Journal of Business*, **53**(3), S13–S29.

McFarland, D. D. (1970). Intragenerational social mobility as a markov process : including a time-stationary markovian model that explains observed declines in mobility rates over time. *American Sociological Review*, **35**(3), 463–476.

McGinnis, R. (1967). A stochastic model of social mobility. *American Sociological Review*, **33**(5), 712–722.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., et Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, **18**, 195–222.

Norris, J. R. (1997). *Markov Chains*. Cambridge University press.

Paradis-Therrien, C. (2007). *BART applied to insurance*. Université de Montréal.

Poissant, M. (2009). *Statistical methods for insurance fraud detection*. Université de Montréal.

Prais, S. J. (1955). Measuring social mobility. *Journal of the Royal Statistical Society*, **118**(1), 56–66.

Spilerman, S. (1972). The analysis of mobility processes by the introduction of independent variables into a markov chain. *American Sociological Review*,

37(3), 277–294.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Annexe A

DENSITÉS CONSIDÉRÉES POUR L'OBTENTION DE LA FONCTION D'AJUSTEMENT

A.1. LOGISTIQUE

Les paramètres sont $\mu \in \mathbb{R}$ et $\sigma > 0$. La densité et la fonction de survie sont données par

$$f(x) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma\left(1+\exp\left(-\frac{x-\mu}{\sigma}\right)\right)^2} \tag{A.1.1}$$

et

$$S(x) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{1+\exp\left(-\frac{x-\mu}{\sigma}\right)} \tag{A.1.2}$$

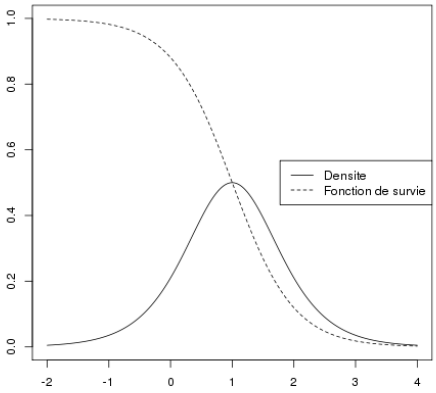


FIG. A.1: Densité et fonction de survie de la loi logistique, $\mu = 1{,}0$, $\sigma = 0{,}5$

A.2. LOGLOGISTIQUE

Les paramètres sont $\alpha > 0$ et $\lambda > 0$. La densité et la fonction de survie, qui sont définies pour $x \geq 0$, sont données par

$$f(x) = \frac{\alpha x^{\alpha-1}\lambda}{(1+\lambda x^{\alpha})^2} \tag{A.2.1}$$

et

$$S(x) = \frac{1}{1+\lambda x^{\alpha}}. \tag{A.2.2}$$

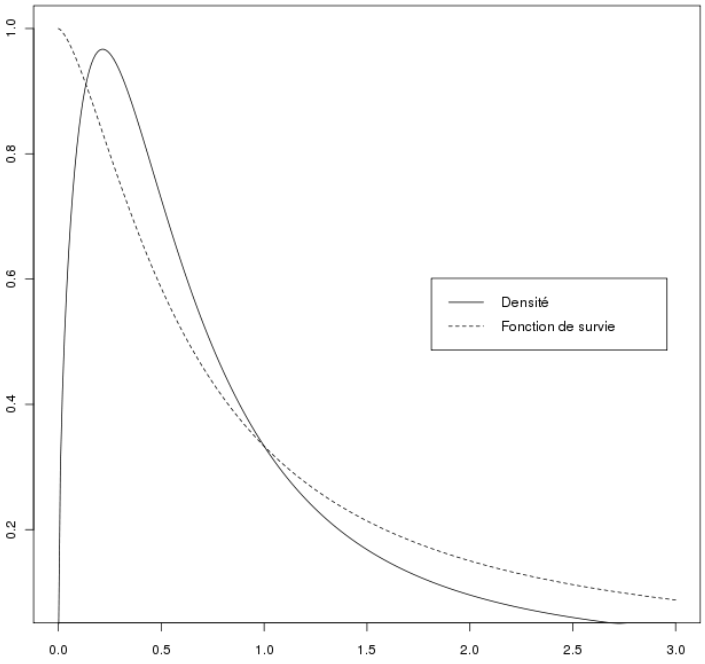


FIG. A.2: Densité et fonction de survie de la loi loglogistique, $\alpha = 1{,}5$, $\lambda = 2{,}0$

A.3. LOGNORMALE

Les paramètres sont $\mu \in \mathbb{R}$ et $\sigma > 0$. La densité et la fonction de survie, qui sont définies pour $x \geq 0$, sont données par

$$f(x) = \frac{\exp\left[-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right]}{x\sqrt{2\pi\sigma^2}} \tag{A.3.1}$$

et

$$S(x) = 1 - \Phi\left(\frac{\log x - \mu}{\sigma}\right) \tag{A.3.2}$$

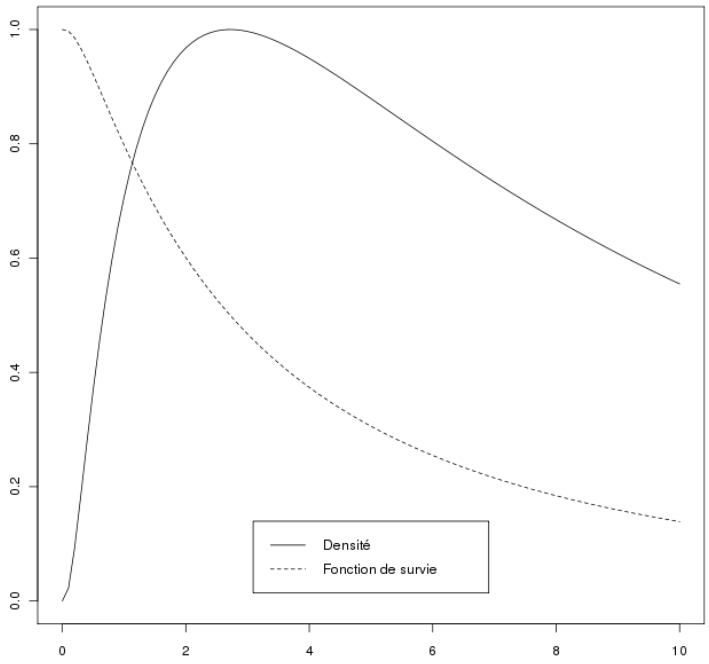


FIG. A.3: Densité et fonction de survie de la loi lognormale, $\mu = 1{,}0$, $\sigma = 1{,}2$

A.4. WEIBULL

Les paramètres sont $\alpha > 0$ et $\lambda > 0$. La densité et la fonction de survie, qui sont définies pour $x \geq 0$, sont données par

$$f(x) = \alpha\lambda x^{\alpha-1} \exp(-\lambda x^{\alpha}) \tag{A.4.1}$$

et

$$S(x) = \exp(-\lambda x^{\alpha}) \tag{A.4.2}$$

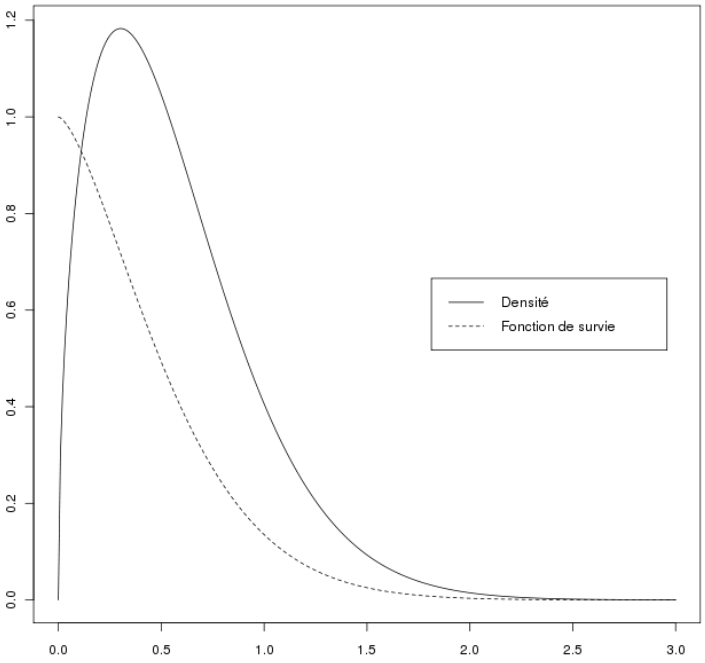


FIG. A.4: Densité et fonction de survie de la loi de Weibull, $\alpha = 1{,}0$, $\lambda = 1{,}2$

A.5. GUMBEL

Les paramètres sont $\mu \in \mathbb{R}$ et $\sigma > 0$. La densité et la fonction de survie sont données par

$$f(x) = \frac{\exp\left(\frac{x-\mu}{\sigma} - \exp\left(\frac{x-\mu}{\sigma}\right)\right)}{\sigma} \tag{A.5.1}$$

et

$$S(x) = \exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right) \tag{A.5.2}$$

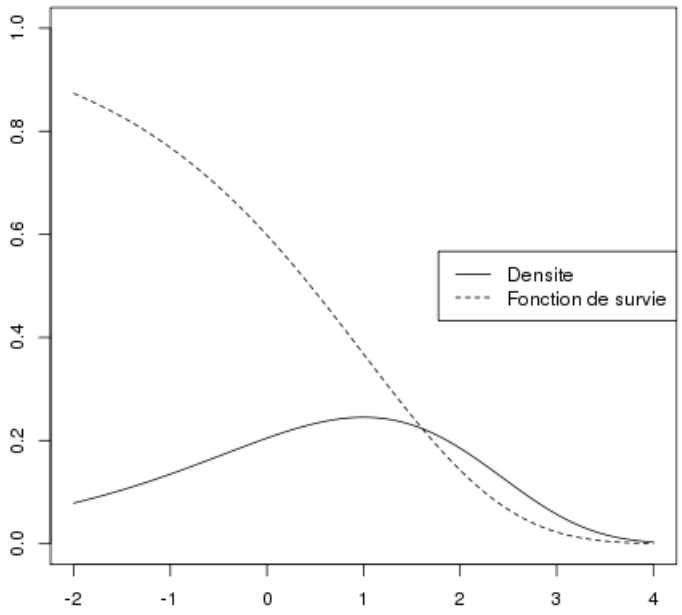


FIG. A.5: Densité et fonction de survie de la loi de Gumbel, $\mu = 1{,}0$, $\sigma = 1{,}5$

A.6. VALEURS EXTRÊMES GÉNÉRALISÉE

Les paramètres sont $\alpha > 0$ (forme), $\mu \in \mathbb{R}$ (location) et $\sigma > 0$ (échelle). La densité et la fonction de survie sont données par

$$f(x) = \frac{\exp\left(\frac{\alpha(x-\mu)}{\sigma} - \exp\left(\frac{x-\mu}{\sigma}\right)\right)}{\sigma\Gamma(\alpha)} \tag{A.6.1}$$

et

$$S(x) = \frac{\Gamma\left(\alpha, \exp\left(\frac{x-\mu}{\sigma}\right)\right)}{\Gamma(\alpha)} \tag{A.6.2}$$

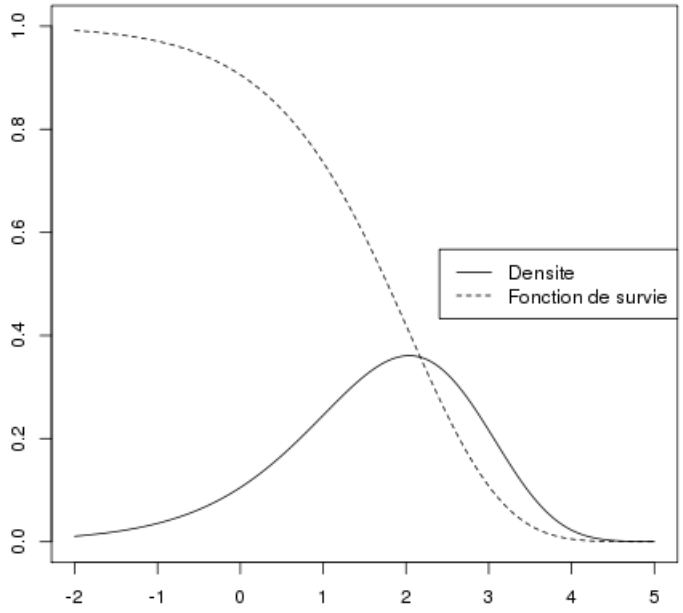


FIG. A.6: Densité et fonction de survie de la loi valeurs extrêmes généralisée, $\alpha = 2{,}0$, $\mu = 1{,}0$ et $\sigma = 1{,}5$