METHODES D'ESTIMATION DE DUREES DE VIE DE CONTRATS D'ASSURANCES AUTOMOBILES

J.M. Marion*, J.M. Loizeau**, A. Oulidi*

*IMA, 44 Rue rabelais, BP 808, 49008 Angers cedex 01
jean-marie.marion@uco.fr
Abder.oulidi@uco.fr

**MAAF, Niort, FRANCE
loizeau@maaf.fr,
http://www.maaf.fr

Résumé. Dans cet article nous étudions le phénomène de résiliation de contrats en utilisant les méthodes classiques (non paramétriques, paramétriques et semi-paramétriques) de durées de vie appliquées à un portefeuille d'une compagnie de taille significative sur le marché français d'assurance non-vie.

1 Introduction

En France, l'assurance automobile est un marché mature avec un faible taux de croissance. De plus, s'agissant d'un secteur convoité, de nouveaux intervenants (banques-assurances, les acteurs de la grande distribution . . .) viennent rejoindre les acteurs traditionnels. Confrontés à une forte concurrence exacerbée par la quasi-stabilité du parc automobile assurable, et face aux mutations importantes de leur référentiel comptable et réglementaire, les assureurs sont désormais incités, plus qu'avant, à développer des modèles optimaux de surveillance et de gestion de leur portefeuille afin, entre autres, de fidéliser les clients les plus rentables et éventuellement de résilier certains contrats. Les analyses d'ordre qualitatif s'avèrent rapidement très insuffisantes, et les assureurs ont recours, de plus en plus, à des techniques statistiques plus élaborées. Des techniques de classification, des modèles de prévisions permettent de réaliser la tarification et d'élaborer des modèles prédictifs de résiliation de certains contrats d'assurance automobile Dans ce papier, on s'intéresse tout particulièrement à l'étude du phénomène de résiliation et à l'élaboration d'un modèle prédictif de durée de vie de contrats Auto. Une application pratique est réalisée sur un extrait du portefeuille d'une compagnie de taille significative sur le marché français d'assurance non-vie. On entend par durée de vie de contrat Auto, la durée séparant la date de résiliation de la date de création de contrat. La date de création de contrat est connue, par contre la date de résiliation n'est pas connue au moment des traitements pour tous les contrats, on dit que les données sont censurées à droite (voir paragraphe 2, pour une définition plus complète). Ces données manquantes compliquent sérieusement l'analyse et nous poussent à utiliser des outils plus adéquats : les modèles de survie (cf Cox D.R. et Oakes D.

(1984), Droesbeke J.J, Fichet B. et Tassi P.(1998),Fermanian J.D. (1993), Li S. (1996)).

En effet, les modèles de survie sont adaptés dès que les phénomènes d'intérêt se modélisent comme des variables aléatoires positives avec éventuellement des données manquantes. Ils sont utilisés dès qu'on cherche à modéliser et à estimer les lois décrivant le temps qui s'écoule entre deux événements à partir d'observations de durées et éventuellement de variables explicatives dites exogènes (ou covariables).

L'analyse des durées de vie peut se faire en utilisant des méthodes d'inférence statistique traditionnelles adaptées aux données censurées (cf Droesbeke J.J et al(1998) ,Fermanian J.D. (1993),Harrington T.R et Fleming D.P. (1991) (modèles paramétriques, semi-paramétriques et non-paramétriques) qui feront l'objet de ce papier et qui seront étudiées en détail au paragraphe suivant. Ces méthodes considèrent que l'on observe des variables aléatoires positives T représentant des durées jusqu'à ce qu'un certain événement ait lieu (résiliation de contrat par exemple).

Il est aussi possible d'aborder les modèles de survie sous forme de processus ponctuels, en considérant les observations comme des processus évoluant au cours du temps. On peut par exemple associer à un contrat un processus " de présence à risque " $Y(t)$ qui vaut 1 à chaque instant t où le contrat est observé et n'a pas encore été résilié 0 sinon, et un processus $N(t)$ qui vaut 1 seulement à partir du moment t où le contrat est résilié 0 sinon. Si le contrat a été résilié au temps aléatoire T, $Y(t)$ vaut 1 tant t que est inférieur à t, puis il vaut 0 sinon, et $N(t)$ prend le relais, valant 0 quand Y vaut 1 et 1 dès que Y vaut 0. Le processus $N(t)$ est appelé processus ponctuel et la connaissance de N suffit à déterminer Y et T. Ces méthodes permettent l'usage de résultats puissants concernant les martingales et les processus prévisibles pour procéder à des estimations et des tests sur les durées de vie. Pour plus de détails, on peut se référer par exemple à (Fermanian J.D. (1993))).

Les modèles de survie, ont été développés pour des applications en biologie, en médecine (biostatistique, épidémiologie . . .), en démographie (espérance de vie aux divers âges . . .), en économie (analyse du marché de travail, durées de vie des entreprises . . .), en finance (défaillances de crédit), en assurance (tables de mortalités d'expériences), en fiabilité (durée de vie de composants industriels), etc. (Li S. (1996),Marubini E. et al (1982), Perrigot R. et al (2004)))

Ce papier est organisé comme suit : la première partie sera consacrée à la présentation détaillée des méthodes statistiques traditionnelles (modèles paramétriques, semi paramétriques et non paramétriques). On donnera aussi brièvement quelques définitions concernant les troncatures et censures. La deuxième partie consistera en une illustration de différentes méthodes à des données réelles extraites du portefeuille d'une compagnie de taille significative sur le marché français d'assurance non-vie.

2 Théorie

Considérons T une variable d'intérêt , c'est à dire une variable aléatoire positive décrivant le temps qui s'écoule entre deux évènements : par exemple la durée de vie d'un contrat qui peut être définie comme la différence entre la date de résiliation et la date de création du contrat.

2.1 Fonction de survie, fonction de hasard

Nous supposerons que la distribution de T possède en tout point une densité de probabilité f, sa fonction de répartition sera notée F.

Les fonctions utilisées habituellement en analyse des données de survie sont

- la fonction de survie S définie par : $S(t) = P(T > t)$;
- la fonction de hasard h définie par : $h(t) = \frac{f(t)}{S(t)}$, qui est une caractéristique locale s'interprétant au point t comme la probabilité instantanée de sortie de l'état sachant que le sujet est encore dans cet état à l'instant t, soit :

$$h(t) = \lim_{\triangle_t \to 0} \frac{1}{\triangle_t} P\left(T \in]t, t + \triangle_t] / T > t\right);$$

- la fonction de hasard intégrée H définie par : $H(t) = \int_0^t h(x)dx$
  notons que :
  $$S(t) = \exp(-H(t)) \text{ puisque } S(0) = 1.$$

Ces fonctions caractérisent entièrement la distribution de T.

2.2 Censures, troncatures

Ce n'est pas toujours T, la variable d'intérêt qui est observée, mais une autre variable C, appelée censure qui donne une information sur T.
Par exemple, nous ne connaissons pas la date de résiliation du contrat : au lieu de T, nous observons (Y, D) avec $y = \inf(T, C)$ et D indicatrice de l'événement $\{T \leq C\}$, si $D = 0$ nous dirons dans ce cas que T est censuré à droite (il existe aussi des censures gauches, dans ce cas, nous observons (Y, D) avec $Y = \sup(T, C)$ et D indicatrice de l'événement $\{T \leq C\}$).
Notons que la censure peut être aléatoire (il est souhaitable que C et T soient indépendantes, pour des problèmes d'identifiabilité de la loi de T (Fermanian J.D. (1993)).
Comme cas particulier, la censure fixe correspond à une variable de censure C dégénérée (ce sera le cas si l'on suppose que C correspond à la date d'aujourd'hui et que l'on considère la durée de vie des contrats jusqu'à cette date, certains contrats ne seront pas encore résiliés, il s'agit alors d'une censure droite fixe).
Il existe aussi d'autres types de censure (censures par intervalle, censure progressive ...)

Il est possible que l'on n'observe T que conditionnellement au fait qu'il appartienne à un ensemble A (on peut étudier la loi de la variable d'intérêt T que si une variable exogène liée à l'individu se trouve appartenir à un certain domaine), on parle alors de troncature (celle-ci est liée à l'échantillon lui-même).

2.3 Les modèles

Trois approches sont possibles pour estimer les durées de vie.

2.3.1 Les modèles non paramétriques

Ces modèles permettent d'éviter le choix à priori d'une loi pour la durée de vie. Nous sommes ramenés dans ce cas à un problème d'estimation fonctionnelle. Parmi les estimateurs de cette classe, nous trouvons l'estimateur de Kaplan-Meier pour la fonction de survie (Kaplan E.L. et Meier P. (1958)).
Considérons un échantillon $(T_1, \ldots, T_n)$ de la variable d'intérêt T, nous observons en réalité $Y_i = \inf(T_i, C_i)$ avec $D_i = \mathbb{I}_{\{}T_i \leq C_i\}$ indicatrice de l'événement $\{T_i \leq C_i\}$ où les $C_i (1 \leq i \leq n)$ désignent les censures droites.
Soit $Y_{(1)}, \ldots, Y_{(n)}$ la statistique d'ordre associée à $Y_1, \ldots, Y_n$ et les indicatrices ordonnées correspondantes, nous définissons $R(t)$ le nombre des " individus à risque " à l'instant t, c'est à dire par exemple les contrats qui sont encore présents à t^- (ni résiliés, ni censurés) et nous notons $M(Y_{(i)})$ le nombre de résiliation à $Y_{(i)}$.
L'estimateur de Kaplan-Meier se définit de façon générale par :

$$\hat{S}(t) = \prod_{\{i, Y_(i) < t\}} \left(1 - \frac{M(Y_{(i)})}{R(Y_{(i)})}\right) \tag{1}$$

Sous certaines conditions, on montre la convergence de cet estimateur vers la fonction de survie et des propriétés de normalité asymptotiques. De l'estimateur de Kaplan-Meier de la fonction de survie, on peut déduire un estimateur de la fonction de hasard cumulé qui n'est autre que :

$$\hat{H}(t) = \sum_{\{i, Y_(i) < t\}} \frac{M(Y_{(i)})}{R(Y_{(i)})} \tag{2}$$

Lorsque n est grand, cet estimateur correspond à l'estimateur de Nelson-Aalen pour la fonction de hasard cumulé.

En ce qui concerne l'estimation non paramétrique de la fonction de hasard on est ramené à des méthodes d'estimation de la densité, on peut alors utiliser des méthodes du type noyau de convolution.
Notons que dans le cas où la population étudiée est fractionnée en plusieurs groupes (liés à des variables exogènes), on utilise des estimateurs non paramétriques pour chacun des sous-groupes.

2.3.2 Les modèles paramétriques

La loi de probabilité de la durée de vie T appartient à une classe de distributions de type connu, fonction de paramètres dont l'objectif sera de les estimer à partir d'un ensemble d'observations (Cox D.R. et Oakes D. (1984)).

Considérons un échantillon $t_1, \ldots, t_n$ issu d'une distribution connue de densité $f(x, \theta)$ où θ est un paramètre inconnu qui peut être vectoriel.
En fait, nous observons $y_1, \ldots, y_n$ où certaines valeurs sont des censures droites, d'autres des censures gauches, enfin certaines valeurs correspondent aux t_i.

Nous noterons : $D_{1_i} = \mathbb{I}_{\{T_i > C_i^d\}}$ où C_i^d correspond à une censure droite, $D_{2_i} = \mathbb{I}_{\{T_i < C_i^g\}}$ où C_i^g correspond à une censure gauche et $D_{3_i} = 1 - D_{1_i} - D_{2_i}$. La fonction de vraisemblance correspondant à l'échantillon $y_1, \ldots, y_n$ s'écrit

$$L(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} \left\{ S(y_i)^{D_{1_i}} F(y_i)^{D_{2_i}} f(y_i)^{D_{3_i}} \right\} \tag{3}$$

Les estimateurs des composantes $\theta_j (1 \leq j \leq p)$ du vecteur θ sont solutions des équations de $\log$ vraisemblance :

$$\frac{\partial \ln L(y_1, \ldots, y_n; \theta)}{\partial \theta_j} = 0 \tag{4}$$

Quand le système des équations ci-dessus est un système d'équations linéaires, les paramètres du vecteur sont déterminés analytiquement. Dans le cas d'un système d'équations non linéaires, une méthode itérative d'optimisation, par exemple la méthode de Newton-Raphson est utilisée pour la détermination des composantes du vecteur.
Les modèles paramétriques permettent l'introduction de variables exogènes ; par exemple on peut penser que la durée de vie d'un contrat d'assurance Auto est liée au Bonus-Malus, au type de contrat, à l'âge du véhicule ... c'est à dire à diverses caractéristiques exogènes dont certaines sont continues.
Nous sommes alors intéressés par les liens entre la variable de durée T et cet ensemble de variables exogènes. Parmi les modèles les plus répandus, nous allons considérer le modèle de régression log linéaire défini par :

$$\ln y_i = \tilde{x}_i \beta + \eta \varepsilon_i \ \text{ pour } \ 1 \leq i \leq n \tag{5}$$

avec :

$\tilde{x}_i$ transposé du vecteur des variables exogènes correspondant à l'individu i

β vecteur des paramètres à estimer

η paramètre d'échelle, par défaut sa valeur est prise égale à 1

ε_i variable aléatoire dont la distribution de probabilité est connue (exponentielle, Weibull, $\log$ logistique ,...)

La $\log$-vraisemblance de l'échantillon est donnée par :

$$\ln L(\zeta_1, \ldots, \zeta_n; \theta) = \sum_{i=1}^{n} \{ D_{1_i} (\ln S_\varepsilon(\zeta_i)) + D_{2_i} (\ln F_\varepsilon(\zeta_i)) + D_{3_i} (\ln f_\varepsilon(\zeta_i)) \} \tag{6}$$

avec $\zeta_i = \frac{\ln y_i - \tilde{x}_i \beta}{\eta}$ et $F_\varepsilon, S_\varepsilon, f_\varepsilon$ désignant respectivement la fonction de répartition, la fonction de survie et la densité de probabilité des ε_i. Les paramètres η, β et ceux de la loi des ε_i sont estimés par la méthode du maximum de vraisemblance puis la méthode de Newton-Raphson. Connaissant la loi des ε_i, on en déduit la fonction de survie de T.

2.3.3 Les modèles semi-paramétriques

Les modèles semi-paramétriques se situent entre les deux approches précédentes. On les utilise lorsque la famille de lois à laquelle appartient la loi de la variable de durée T n'est pas

totalement spécifiée, on cherche de plus à évaluer l'effet de variables exogènes sur T.
Parmi ces modèles, le plus utilisé est le modèle à hasard proportionnel de Cox (Cox D.R. (1972)). La fonction de hasard, s'écrit dans ce cas :

$$h(t/x) = h_0(t)\exp(\tilde{x}\beta) \tag{7}$$

où

h_0 est la fonction de hasard de base (non spécifiée) - elle peut s'estimer non paramétriquement,

$\tilde{x}$ le transposé du vecteur des variables exogènes,

β est le vecteur des paramètres à estimer.

Ce modèle peut se généraliser en considérant des variables exogènes dépendant du temps.

Remarquons que h est séparable en deux termes dont l'un h_0 est une fonction du temps et l'autre ne dépend que des variables exogènes. De cette remarque, nous déduisons que :

$$\frac{h(t/x_1)}{h(t/x_2)} = \exp(\tilde{x}_1\beta - \tilde{x}_2\beta) \tag{8}$$

c'est à dire que le rapport des fonctions de hasard pour deux individus de caractéristiques x_1 et x_2 ne dépend pas de t, mais seulement de x_1 et x_2 (ce rapport est dit " hazard ratio ").
Si l'on note S_0 la fonction de survie de base associée à h_0 , on a la relation suivante :

$$S(t/x) = [S_0(t)]^{\exp(\tilde{x}\beta)}, \tag{9}$$

ce qui permet d'obtenir une estimation de S connaissant l'estimation du vecteur β.
Pour estimer les composantes du vecteur β, à partir d'un échantillon ordonné $(y_{(1)}, \ldots, y_{(n)})$, on calcule la fonction de vraisemblance partielle de Cox qui n'est autre que (s'il n'y a pas de données censurées) :

$$L(y_{(1)}, \ldots, y_{(n)}; \beta) = \prod_{i=1}^{n} \frac{\exp(\tilde{x}_i\beta)}{\sum_{k\in R(y_{(i)})}\exp(\tilde{x}_k\beta)} \tag{10}$$

Les estimateurs des composantes β_j du vecteur β sont solutions des équations de $\log$ vraisemblance :

$$\frac{\partial \ln L(y_{(1)}, \ldots, y_{(n)}; \beta)}{\partial \beta_j} = \sum_{i=1}^{n}\left(x_{ij} - \frac{\sum_{k\in R(y_{(i)})} x_{kj}\exp(\tilde{x}_k\beta)}{\sum_{k\in R(y_{(i)})}\exp(\tilde{x}_k\beta)}\right) \text{ pour } 1 \le j \le q \tag{11}$$

où q désigne le nombre de composantes du vecteur β. x_{ij} est la $j^{\text{ème}}$ composante du vecteur des variables exogènes pour l'individu i. Les solutions en β_j sont calculées par la méthode de Newton-Raphson. Le test du score, par exemple, permet de tester la significativité des composantes de β.
On montre que la présence de la censure à droite ne modifie pas la valeur de $L(y_{(1)}, \ldots, y_{(n)}; \beta)$, l'ensemble des éléments à risque étant connu à chaque instant. Par conséquent, l'introduction de données censurées à droite ne change pas la valeur de l'estimateur $\hat{\beta}$ de β.

Notons que dans ce cas, une estimation de la fonction de survie peut être donnée par la relation suivante :

$$\hat{S}(t/x) = \prod_{\{j;y(j)<t\}} \hat{\mathcal{V}}_j^{\exp(\tilde{x}\hat{\beta})} \tag{12}$$

où les $\hat{\mathcal{V}}_j$ sont solutions des équations de vraisemblances :

$$\sum_{l \in D_{3_j}} \frac{\exp(\tilde{x}_l\hat{\beta})}{1-\hat{\mathcal{V}}_j^{\exp(\tilde{x}_l\hat{\beta})}} = \sum_{l \in R(y(l))} \exp(\tilde{x}\hat{\beta}) \text{ pour } 1 \leq j \leq z \tag{13}$$

dans lesquelles z est le nombre de durées de vie distinctes et D_{3_j} correspond aux durées de vie effectivement observées dans l'échantillon.

3 Application à des données de contrats d'assurances automobiles

Dans ce paragraphe seront présentées les données étudiées ainsi que l'illustration de méthodes présentées précédemment.

3.1 Données

Les données analysées sont issues d'un portefeuille provenant d'une compagnie de taille significative sur le marché français d'assurance non-vie. La sélection concerne uniquement le portefeuille Auto géré par une agence, ce qui correspond à un territoire géographique limité à quelques communes. Pour des raisons de confidentialité, cet extrait est non représentatif du portefeuille global. Néanmoins, cette sélection conserve des caractéristiques suffisamment pertinentes pour permettre une application valable des méthodes d'estimation de durées de vie et une interprétation des résultats de cette agence. Après avoir éliminé quelques valeurs aberrantes [1], le fichier final que nous étudions comporte 1557 contrats Autos.
Dans ce papier, l'objectif étant de présenter des méthodes pour estimer la durée de vie des contrats d'assurance Auto, nous avons choisi une agence pour laquelle le pourcentage des résiliations autres qu'en provenance des assurés était négligeable. Tous les types de résiliations de contrats Autos ont donc été pris en compte, qu'elles soient à l'initiative du client ou de la compagnie (suite à disparition du risque, vente ou fin de vie du véhicule suite panne ou accident, pour non-paiement, ou encore d'un commun accord entre assureur et assuré ...). Bien entendu, ces différentes résiliations ne sont pas homogènes et peuvent avoir des résultats sensiblement éloignés en terme de durée de vie du contrat pour un portefeuille dans lequel existe toutes les formes de résiliation.
Ces contrats ont été crées entre le 13 juin 1974 et le 28 décembre 1995, leur date de résiliation est située après le 1er janvier 1996. Dans l'étude qui a été faite, la variable d'intérêt est la durée de vie des contrats (notée DurVie), c'est à dire que si la date de résiliation est située avant le 31 décembre 2000 nous avons considéré la différence entre la date de résiliation et la date de

[1] Les valeurs aberrantes concernent uniquement quelques dizaines de contrats pour lesquels la variable date de mise en circulation du véhicule n'est pas exploitable.

création sinon nous avons considéré l'écart entre le 31 décembre 2000 et la date de création du contrat (nous considérons donc une censure droite fixe).
Les variables exogènes considérées sont observées le 1er janvier 1996, il s'agit de :

- L'âge du véhicule (notée AgeVehic), c'est une variable quantitative définie comme l'écart entre le 1er janvier 1996 et la date de mise en circulation du véhicule. (pour certaines illustrations, cette variable a été codée sous le non AgeVehicCode de la sorte :
- AgeVehicCode1 correspond à un code AgeVehic inférieur ou égal à un 1 ;
- AgeVehicCode2 correspond à un code AgeVehic compris entre 1 et 4 (inclus) ;
- AgeVehicCode3 correspond à un code AgeVehic compris entre 4 et 8 ;
- AgeVehicCode4 correspond à un code AgeVehic strictement supérieur à 8) ;
- Le Bonus-Malus (noté CodeMalus), cette variable a été regroupée en trois classes :
  - CodeMalus1 correspond à un code Bonus-Malus $= 0,5(50\%$de bonus$)$,
  - CodeMalus2 correspond à un code Bonus-Malus compris entre $0,5$ et $0,7$ (entre 30% et 50% de bonus),
  - CodeMalus3 correspond à un code Bonus-Malus strictement supérieur à $0,7$ (moins de 30% de bonus, ou bien malus) ;
- La formule d'assurance (notée Code), cette variable a été regroupée en trois classes :
  - Code1 correspond à Tierce Intégrale (formule complète Tous risques),
  - Code2 correspond à Tiers Maxi ou Tierce Collision (formule RC+ dommages mais hors Tous risques),
  - Code3 correspond à Tiers Simple (formule RC seule).

3.2 Statistique exploratoire

Une étude statistique descriptive montre les répartitions suivantes :

Contrats	Résiliés	Censurés	Total
CodeMalus1	483	139	622
CodeMalus2	270	149	419
CodeMalus3	231	285	516
Total	**984**	**573**	**1557**

Contrats	Résiliés	Censurés	Total
Code1	324	190	514
Code2	443	202	645
Code3	219	181	398
Total	**984**	**573**	**1557**

Contrats	Résiliés	Censurés	Total
AgeVehicCode1	66	61	127
AgeVehicCode2	245	137	382
AgeVehicCode3	295	193	488
AgeVehicCode4	378	182	560
Total	**984**	**573**	**1557**

Le portefeuille étudié est principalement constitué de contrats avec 50% de bonus, de formules RC et dommages hors tous risques et de véhicules âgés de plus de 4 ans. Du fait de la sélection

	Moyenne			BorneInf		
	Tout	Résiliés	Censurés	Tout	Résiliés	Censurés
Global	9.47	10.38	7.89	9.27	10.13	7.62
CodeMalus1	10.66	11.12	9.08	10.34	10.77	8.36
CodeMalus2	9.32	9.86	8.33	8.94	9.38	7.77
CodeMalus3	8.14	9.44	7.08	7.86	8.93	6.84
Code1	9.61	10.39	8.28	9.26	9.95	7.77
Code2	10	10.77	8.32	9.70	10.41	7.85
Code3	8.40	9.55	7.01	8.04	9.02	6.65
AgeVehicCode1	9.14	10.68	7.48	8.42	9.61	6.68
AgeVehicCode2	8.95	9.55	7.88	8.55	9.02	7.33
AgeVehicCode3	9.44	10.48	7.86	9.11	10.03	7.42
AgeVehicCode4	9.90	10.79	8.08	9.58	10.40	7.57

des données, ces éléments ne peuvent pas être comparés à ceux du portefeuille global de la compagnie ou à ceux du marché. Le tableau qui suit donne quelques statistiques concernant la variable d'intérêt DurVie. On donne la durée de vie moyenne puis

$$\text{BorneInf} = \overline{x} - \frac{2\sigma}{\sqrt{n}} \text{ et } \text{BorneSup} = \overline{x} + \frac{2\sigma}{\sqrt{n}} \tag{14}$$

où n représente la taille de l'échantillon considéré. On donne aussi la durée de vie minimale et maximale pour toutes les données et pour chacune des variables exogènes considérées.

	BorneSup			Min			Max		
	Tout	Résiliés	Censurés	Tout	résiliés	Censurés	Tout	Résiliés	Censurés
Global	9.66	10.63	8.16	0.14	0.14	5	21.78	21.78	19
CodeMalus1	10.98	11.46	9.80	0.59	0.59	5	20.88	20.88	19
CodeMalus2	9.69	10.34	8.89	0.14	0.14	5	21.78	21.78	18
CodeMalus3	8.42	9.95	7.33	0.77	0.77	5	21.76	21.76	16
Code1	9.96	10.83	8.78	0.59	0.59	5	20.10	20.10	18.92
Code2	10.31	11.14	8.79	0.14	0.14	5	21.78	21.78	19
Code3	8.76	10.09	7.37	0.77	0.77	5	19.67	19.67	18.33
AgeVehicCode1	9.87	11.75	8.28	0.59	0.60	5	20.10	20.10	16.44
AgeVehicCode2	9.35	10.08	8.42	0.14	0.14	5	18	17.35	18
AgeVehicCode3	9.78	10.92	8.30	1.29	1.29	5	21.78	21.78	18.92
AgeVehicCode4	10.23	11.18	8.59	0.56	0.56	5	21.76	21.76	19

Ces statistiques nous indiquent une durée de vie moyenne des contrats de $9,5$ ans au global sur ce portefeuille, et des différences plus ou moins importantes entre les segments étudiés de $8,1$ à $10,7$ ans. Par contre, ces statistiques simples ne nous indiquent pas si les écarts sont significatifs et nous donnent peu d'information sur les distributions des durées de vie ou fonctions de survie. C'est ce que proposent les méthodes suivantes d'estimation des durées de vie.

3.3 Estimation de Kaplan-Meier

Nous représentons les fonctions de survie en utilisant l'estimateur de Kaplan-Meier.

Sur le graphique (Fig. 1), nous constatons une différence entre les trois " CodeMalus ". La courbe correspondant au CodeMalus1 suggère que globalement ce groupe a davantage de chances pour que la durée de vie des contrats soit plus grande que pour les deux autres groupes.

La statistique du Log-rank prenant la valeur 12.2, confirme une différence significative entre les trois courbes avec une p-value de $0,002$ environ.

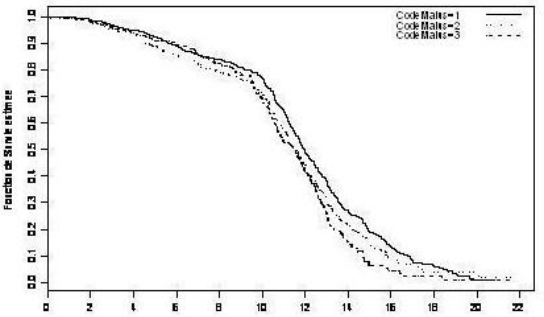

FIG. 1 – *Durée de vie des contrats*

Le graphique (Fig. 2) présente une estimation des fonctions de survie correspondant aux trois " Code ", c'est à dire aux différentes formules d'assurance. Là encore, la statistique du Log-rank prenant la valeur $8,8$ confirme une différence significative entre les groupes liés aux formules d'assurance avec une p-value de $0,012$ environ. La courbe correspondant au Code3, c'est à dire à la formule Tiers Simple, laisse à penser que ce groupe a plus de chances que la durée de vie des contrats soit plus faible que pour les deux autres groupes.

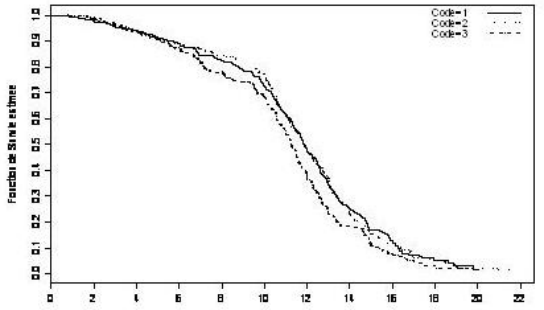

FIG. 2 – *Durée de vie des contrats*

Puisque les courbes sont significativement différentes, cela signifie que les variables exogènes jouent un rôle dans la durée de vie des contrats.
Le graphique (Fig. 3) qui suit présente une estimation des fonctions de survie toujours par la méthode de Kaplan-Meier, mais ici en tenant compte du découpage de la variable AgeVehic (âge du véhicule) en 4 classes comme indiqué dans l'introduction. Remarquons que la statistique du Log-rank prend la valeur $7,7$ avec une p-value de 0.052 environ.

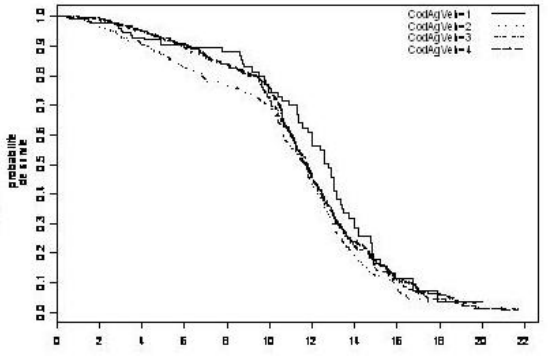

FIG. 3 – *Durée de vie des contrats*

3.4 Estimation paramétrique

Pour la recherche d'un modèle de survie paramétrique, nous avons testé plusieurs lois : la loi exponentielle, la loi de Weibull, la loi log-logistique, la loi log-normale, la loi des valeurs extrêmes . . . En utilisant le critère de la vraisemblance maximale, la loi log-logistique est apparue la mieux adaptée pour la modélisation de la durée de vie des contrats Auto considérés. La figure suivante montre un exemple de courbe de survie obtenue pour Code $= 2$, CodeMalus $= 2$, AgeVehic $= 7$. Le vecteur $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ où$\hat{\beta}_1$ correspond à l'estimation du coefficient β_1 associé à la variable Code, $\hat{\beta}_2$ correspond à l'estimation du coefficient β_2 associé à la variable CodeMalus et $\hat{\beta}_3$ à l'estimation du coefficient β_3 associé à la variable AgeVehic a pour valeurs :$\hat{\beta}_1 = -0.0739$ (avec une p-value de 0.001) ; $\hat{\beta}_2 = -0.0191$ (avec une p-value de 0.022) ;$\hat{\beta}_3 = 0.0101$ (avec une p-value de 0.001). La loi log-logistique étant définie par sa densité

$$f(t) = \frac{\lambda\alpha(\lambda t)^{\alpha-1}}{(1+(\lambda t)^{\alpha})^2} \tag{15}$$

avec comme estimation des paramètres :$\hat{\alpha} = 3.731$; $\hat{\lambda} = 0.089$

3.5 Estimation par le modèle de Cox

Nous avons utilisé le modèle de Cox en stratifiant sur une variable exogène. Dans un premier temps, nous avons stratifié sur la variable CodeMalus. Le coefficient β_1 associé à la variable AgeVehic est estimé par $\hat{\beta}_1 = -0.0265$ (la p-value associée est de 0.00540), le coefficient β_2 associé à la variable Code est estimé par $\hat{\beta}_2 = 0.2244$ (la p-value associée est de 0.00058).
Le graphique (Fig. 4) correspond à l'estimation de la fonction de survie correspondant à : Code $= 2$; AgeVehic $= 7$.
Nous avons ensuite utilisé un modèle de Cox en stratifiant sur la variable exogène Code. Le coefficient β_1 associé à la variable AgeVehic est estimé par $\hat{\beta}_1 = -0.025$ (la p-value associée

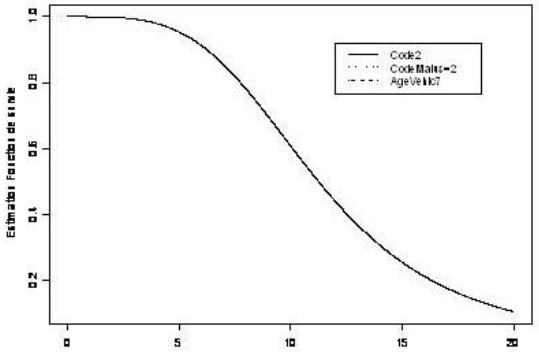

FIG. 4 – *Durée de vie des contrats*

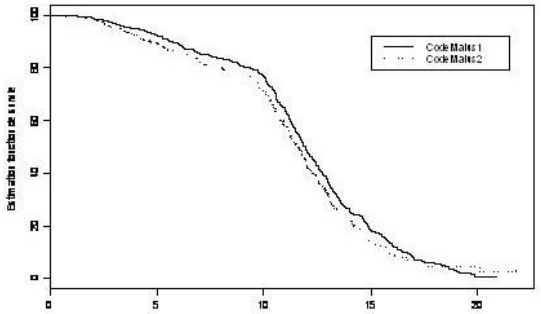

FIG. 5 – *Durée de vie des contrats*

est de 0.041), le coefficient β_2 associé à la variable CodeMalus est estimé par $\hat{\beta}_2 = 0.128$ (la p-value associée est de 0.021). Le graphique (Fig. 4) correspond à l'estimation de la fonction de survie correspondant à : CodeMalus $= 2$; AgeVehic $= 7$.

4 Conclusion

L'estimation des durées de vie de contrats Auto répond à de nombreux enjeux. Elles peuvent être utilisées pour évaluer l'amortissement des coûts d'acquisition d'un contrat, mais aussi pour calculer la rentabilité d'un contrat tenant compte de sa durée de vie prévue (on obtient ainsi un des éléments de la valeur du client), et enfin tous ces éléments peuvent être analysés de manière segmentée en fonction de critères pour affiner le ciblage marketing, les tarifs. . .

Cette étude sur le phénomène de résiliation de contrats a permis d'illustrer les méthodes classiques d'estimation des durées de vie appliquées à un portefeuille d'une compagnie d'assurance non-vie.

En l'absence d'information à priori sur la forme de la fonction de survie, nous l'avons estimé d'abord par la méthode non-paramétrique de Kaplan-Meier. Puis, pour introduire des variables exogènes dans le modèle, nous avons étudié des méthodes paramétriques (modèles de régression log-linéaire) ou la loi log-logistique est apparue la mieux adaptée à nos données pour la modélisation de la durée de vie des contrats Auto. Enfin un modèle semi-paramétrique de Cox a été utilisé en stratifiant sur différentes variables exogènes lorsque les hypothèses liées au modèle le permettaient. Des compléments pourraient être apportés, en particulier dans les méthodes d'estimation semi-paramétriques : il serait possible d'étudier le modèle de Cox avec des variables dépendant du temps ce qui permettrait d'avoir des modèles en meilleure adéquation avec l'ensemble des données observées. Des techniques de comparaisons des différentes méthodes pourraient aussi être utilisées par exemple en considérant un échantillon d'apprentissage et un fichier test qui permettrait de préciser le choix de la méthode en fonction du portefeuille étudié.

Références

ANDERSEN, P.K., BORGAN, Ø., GILL, R.D., et KEIDING, N.,(1993). Statistical models base on counting processes. *Springer Verlag.*

BOSQ D. et LECOUTRE J.P., (1987). Théorie de l'estimation fonctionnelle. *Economica.*

COX D.R.(1972). Regression models and life-tables, *J. Roy.Statist. Soc.B* 34, pp 187- 220.

COX D.R. et OAKES D. (1984). Analysis of survival data. *Edition Chapman and Hall.*

DROESBEKE J.J, FICHET B, TASSI P.,éditeurs (1989). Analyse statistique des durées de vie : Modélisation et données censurées. *Economica.*

FERMANIAN J.D,(1993). Modèles de durées, Cours ENSAE- Bibilothèque de l'ENSAE.

HARRINGTON,T.R., et FLEMING, D.P.(1991). Counting processes and survival analysis. *Wiley.*

KALBFLEISH J.D. et PRENTICE R.L. (1980), The statistical analysis of failure time data. *New York : Wiley.*

KAPLAN E.L. et MEIER P., (1958). Non parametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53, pp 457-481.

LAWLESS J.F., (1982). Statistical Models and Methods for lifetime Data. *Wiley*.

LI, S.,(1996). Survival analysis. *Marketing Research*, 7(4), 17-23.

MARUBINI E et VALSECCHI G.M.,(1982). Analysing Survival Data from Clinical Trials and Observational Studies. Wiley, Chichester.

PERRIGOT R., CLIQUET G., MESBAH M. ,(2004). Possible applications of survival analysis in franchising research, Int. *Rev. of Retail, Distribution and Consumer Research,* Vol.14, N.1, 129-143, January 2004.

Summary

In this paper we study cancellation of contract phenomenon by using the traditional methods (nonparametric, parametric and semi-parametric) of life time applied to a portfolio of a significant size company to the French market of non-life insurance.