Université Claude Bernard – Lyon 1
INSTITUT DE SCIENCE FINANCIERE ET D'ASSURANCES

Mémoire présenté devant l'Institut de Science Financière et d'Assurances pour l'obtention du diplôme d'Actuaire de l'Université de Lyon

le 11 juillet 2011

Par : Xavier Milhaud
Titre : Segmentation et modélisation des comportements de rachat en Assurance Vie

Confidentialité : Oui (2 ans)

Membre du Jury I.A. :
M. Pierre THEROND

Membre du Jury I.S.F.A. :
M. Jean Claude AUGROS
M. Alexis BIENVENÜE
M. Areski COUSIN
Mme Diana DOROBANTU
Mme Anne EYRAUD-LOISEL
M. Stéphane LOISEL
Melle Esterina MASIELLO
Mme Véronique MAUME-DESCHAMPS
M. Frédéric PLANCHET
M. François QUITTARD-PINON
Mme Béatrice REY-FOURNIER
M. Christian-Yann ROBERT
M. Didier RULLIERE

Secrétariat :
Mme Marie-Claude MOUCHON

Bibliothèque :
Mme Michèle SONNIER

Entreprise :
AXA Global Life

Directeur de Mémoire en entreprise :
M. Vincent LEPEZ

Invité : M. Edouard DEBONNEUIL

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

Signature du responsable entreprise

Signature du candidat

50 avenue Tony Garnier - 69366 LYON Cedex 07

Abstract

Keywords : *surrender behaviour*, *heterogeneity*, *classification*, *generalized linear models*

Insurers have been concerned about surrenders for a long time especially in Saving business, where huge sums are at stake. The emergence of the European directive Solvency II, which promotes the development of internal risk models (among which a complete unit is dedicated to surrender risk management), strengthens the necessity to deeply study and understand this risk. In this memoir we investigate the topics of segmenting and modeling surrenders in order to better know and take into account the main risk factors impacting policyholders' decisions. We find that several complex aspects must be specifically dealt with to predict surrenders, in particular the heterogeneity of behaviours and their correlations as well as the context faced by the insured. Combining them, we develop a methodology that seems to provide good results on given business lines, and that moreover can be adapted for other products with little effort.

Résumé

Mots-clés : *comportement de rachat*, *hétérogénéité*, *classification*, *modèles linéaires généralisés.*

La question du rachat préoccupe les assureurs depuis longtemps notamment dans le contexte des contrats d'épargne en Assurance-Vie, pour lesquels des sommes colossales sont en jeu. L'émergence de la directive européenne Solvabilité II, qui préconise le développement de modèles internes (dont un module entier est dédié à la gestion des risques de comportement de rachat), vient renforcer la nécessité d'approfondir la connaissance et la compréhension de ce risque. C'est à ce titre que nous abordons dans ce mémoire les problématiques de segmentation et de modélisation des rachats, dans le but de mieux connaître et prendre en compte l'ensemble des facteurs-clefs qui jouent sur les décisions des assurés. L'hétérogénéité des comportements et leur corrélation ainsi que l'environnement auquel sont soumis les assurés sont autant de difficultés à traiter de manière spécifique afin d'effectuer des prévisions. Nous développons ainsi une méthodologie qui, sur certaines lignes d'affaires, aboutit à des résultats très encourageants ; et qui a l'avantage d'être réplicable en l'adaptant aux spécificités d'autres produits.

Acknowledgements

Je remercie bien évidemment mes encadrants de mémoire, Stéphane Loisel et Véronique Maume-Deschamps, pour leur oeil averti sur les problématiques du rachat ainsi que pour leur apport scientifique essentiel. Leurs compétences techniques de haute volée combinées à leurs connaissances multiples et variées du monde de l'Assurance dans sa globalité ont été un des facteurs clefs du succès de ce mémoire. Je tiens également à remercier Marie-Pierre Gonon qui a contribué à l'élaboration de ce mémoire par l'intermédiaire d'un article que nous avons co-écrit.

En deuxième lieu, j'aimerais également souligner tout le soutien de Vincent Lepez, Edouard Debonneuil et plus généralement de l'entreprise AXA Global Life dans la conduite de cette étude et dans la rédaction de ce mémoire. Tous mes collègues ont joué leur rôle et ont apporté leur pierre à l'édifice, à travers des discussions autour d'un café ou d'autres boissons "énergisantes". J'ai complètement conscience que le bon déroulement de cette étude ne se serait pas fait sans eux.

Enfin je dédie ce mémoire à ma famille et à Marion qui se reconnaîtra. Merci pour ton soutien et ta compréhension de mon rythme pas toujours académique (surtout ces derniers temps !), bien que je fasse le maximum d'effort pour ne pas trop perturber tout le petit monde qui m'entoure ! Je remercie aussi mes frères et parents pour tout (y compris leur vin, mais pas que... !), mes grand-parents et cousins, la famille Valette pour énormément de choses, Agos qui m'a beaucoup fait rire, ma barque qui m'a apporté beaucoup de bonheur et de moments tranquilles pour méditer, la carpe tarnaise, et tous les amis de la "banlieue" de Marssac-sur-Tarn. Les autres que j'aurai oublié ne m'en voudront pas !

Table des matières

Introduction

Le contrat d'assurance vie est un accord entre une compagnie d'assurances qui prend l'engagement irrévocable de verser des prestations au bénéficiaire du contrat en fonction de la réalisation d'événements aléatoires viagers, en échange de quoi le souscripteur prend l'engagement **révocable** de verser des cotisations en fonction de la réalisation d'événements viagers. Le risque de rachat est omniprésent dans les problématiques de valorisation et de provisionnement des contrats d'épargne dans les sociétés d'assurance vie. Pour satisfaire par exemple à un besoin de liquidité immédiat, l'assuré peut à tout moment résilier son contrat et récupérer tout (rachat total) ou partie (rachat partiel) de son épargne capitalisée, éventuellement diminuée de pénalités prévues à cet effet et dépendantes des conditions fixées lors de la souscription. Une bonne compréhension du rachat et de ses facteurs explicatifs (Milhaud et al. (2011)), et plus globalement du comportement des assurés permet :

- d'adapter les caractéristiques et clauses lors de la création de nouveaux produits, avec pour objectif la rétention de clients, le gain de parts de marché ;
- de mettre en place de meilleures stratégies de gestion actif-passif.

Deux questions sous-jacentes au rachat doivent être abordées : tout d'abord les conséquences financières d'un mauvais choix de modélisation pour les lois de rachat et ensuite son impact sur les garanties. Nous nous focalisons dans ce mémoire sur le premier point, étroitement lié au contexte économique et financier et donc à la dynamique des taux d'intérêt. Selon la position de l'assureur (phase d'investissement ou de désinvestissement) et son anticipation du marché, un scénario haussier aussi bien que baissier des taux d'intérêts peut avoir des conséquences importantes au niveau de sa gestion actif-passif et de son stock de réserves. Ces conséquences peuvent même devenir critiques en cas de rachat massif (dans un scénario haussier) ou d'absence de rachat (scénario baissier), obligeant ainsi l'assureur à emprunter ou à verser un taux garanti supérieur au rendement de ses propres actifs. Nous distinguons ainsi qu'un problème d'adéquation se pose pour l'assureur dans tous les cas de figure.

Habituellement, l'assureur fait l'hypothèse que son portefeuille d'assurés est composé de personnes se comportant indépendamment les unes des autres, ce qui est relativement juste en régime de croisière. Néanmoins, un problème majeur se pose dans le cas d'une perturbation de l'équilibre économique et financier : cette hypothèse est alors clairement inadaptée et les comportements des assurés peuvent devenir fortement corrélés (Milhaud et al. (2010)). La recherche académique s'est penchée sur le sujet et les travaux de Lee et al. (2008) apparaissent comme un premier essai de modélisation dynamique du comportement humain, de même que ceux de Fum et al. (2007), Kim et al. (2008) et Pan et al. (2006) qui étudient les réactions humaines en cas de panique. D'un point de vue plus quantitatif, Loisel & Milhaud (2011) présentent certains outils théoriques servant à modéliser l'interaction et la corrélation entre les comportements des assurés, de même que McNeil et al. (2005) qui posent la question dans un contexte plus général.

Des problématiques telles que l'anti-sélection (Bluhm (1982)) et l'aléa moral ont aussi une importance toute particulière, notamment dans un contexte de contrat d'assurance vie prévoyance où la santé des assurés est la question centrale dans le processus de tarification et de gestion des risques. Il est par exemple interdit en France de racheter des contrats de type rente viagère pour éviter le phénomène d'anti-sélection. En épargne, la santé des marchés financiers a une incidence directe sur le comportement de rachat des assurés, créant une corrélation entre leurs décisions. Vandaele & Vanmaele (2008), Bacinello (2005), Kuen (2005) et Tsai et al. (2002) entre autres ont développé des méthodes de valorisation financière de l'option de rachat. Ces méthodes ont vocation à être améliorées pour une meilleure prise en compte de la modélisation comportementale et des réelles questions et besoins de l'assuré lors du rachat, notamment en ce qui concerne la rationalité de son choix.

Nous dressons dans la suite de cette introduction un bref panorama du rachat en France et évoquons certains aspects-clefs dans la compréhension de ce risque. Ensuite, nous abordons la question du risque de rachat et de son évaluation dans le contexte réglementaire de la directive européenne Solvabilité II.

L'assurance vie reste aujourd'hui le placement préféré des Français en offrant la meilleure combinaison risque-rendement-fiscalité : fin 2010, plus de 27 millions de contrats avaient déjà été ouverts et les prestations annuelles versées par les organismes d'assurance vie dépassaient 106 * milliards d'euros ! Les offres sont multiples et permettent de répondre aux besoins des épargnants, les principaux critères étant la liquidité des sommes investies, le rendement et la sécurité financière associée aux entreprises d'assurances. La liquidité se traduit principalement, pour les contrats d'épargne monosupports (un seul support en euros, à valeur plancher garantie) et multisupports, par la liberté de sortir du contrat sans perdre les avantages du produit. Pour cela, l'assuré devra connaître quelques bases de fiscalité applicables aux contrats d'assurance vie. Cette fiscalité évolue régulièrement mais continue d'offrir certains avantages, notamment pour les contrats d'une durée de détention supérieure à huit ans. Cette caractéristique est clairement identifiable dans les observations passées, avec un pic de rachat lors de la neuvième année de présence en portefeuille. L'assuré devra aussi porter attention aux pénalités qui pourraient lui être prélevées au moment de la sortie, qui, selon les codes des assurances, sont cependant limitées à 5% des intérêts de la somme épargnée sur dix ans. Enfin, il pourra étudier les options qui lui sont offertes dans le contrat pour bénéficier de cette liquidité sans pour autant clore son contrat d'assurance vie et ainsi ne pas perdre son antériorité fiscale, au travers notamment des possibilités de rachat partiel ou d'avance. On entend par rachat partiel la possibilité qui est offerte à un assuré de ne racheter qu'une partie de son épargne, et ainsi ne pas demander la fermeture de son contrat. L'avance quant à elle est assimilable à un prêt que consent l'assureur envers l'assuré, l'épargne de l'assuré constituant alors le collatéral. D'autres clients, souvent plus fortunés, sont également intéressés par le caractère non rachetable de leur contrat, mais cela est un autre sujet. Le phénomène de rachat est important pour les assureurs français. Meilleure en est leur connaissance, meilleure sera leur anticipation de gestion des flux entrants et sortants en termes de trésorerie. Ils pourront même parfois en profiter pour améliorer leur offre produit. Alors, quelles sont les variables explicatives du choix de l'assuré, comment pouvons-nous projeter les comportements humains, pouvons-nous répondre en moyenne ? Tout d'abord en termes de variables explicatives, on a observé pendant de nombreuses années l'influence directe de la fiscalité et des

∗. Source : Fédération Française des Sociétés d'Assurances (FFSA)

profils patrimoniaux des assurés. Ces observations avaient pour avantage qu'elles étaient connues parfaitement de l'assureur et pouvaient ainsi faire l'objet de modélisations relativement adaptées.

Puis, on a introduit des effets moins évidents tels que la tenue des marchés financiers. Lorsque les taux baissent, les assurés seraient-ils plus fidèles ? Lorsque les taux montent, auraient-ils tendance à sortir ? Rien n'est moins évident, d'autant que les données selon lesquelles les variations de taux pourraient influencer le comportement des assurés ne sont pas nombreuses. Même au plus haut de la crise financière que l'on vient de connaître, la fidélité des clients n'a pas été énormément entamée. Ce n'est pas tant que les assurés ne suivent pas l'actualité ou ne prennent pas le temps de modifier leur allocation d'épargne, mais plutôt qu'il leur est difficile d'évaluer à quel taux de marché il serait intéressant pour eux d'aller souscrire ailleurs en tenant compte de nouveaux frais d'entrée, d'une perte de l'antériorité fiscale, etc. Ces nombreux facteurs compliquent les décisions à prendre de manière rationnelle. En revanche, c'est lorsque l'on commence à s'intéresser au rendement relatif des contrats d'assurance vie entre eux que l'épargnant devient plus vigilant, voire plus susceptible. Et c'est ainsi que les assureurs se sont penchés sur cette question. Combien de clients vont racheter leur contrat lorsque je servirai un taux de rendement inférieur à celui de mon concurrent, et à partir de quel écart de taux commenceront-ils à réagir ? Là encore, les statistiques ne sont pas nombreuses. Quelques assureurs ont démarché leurs clients pour connaître leur sensibilité à un écart de taux, mais les résultats obtenus ne permettent pas de modéliser de façon fiable une courbe de comportement. D'autres facteurs, comme le relais d'information par la presse quant aux taux garantis, peuvent influencer les décisions des souscripteurs. A ce titre, de nombreux assureurs ont élu un produit vitrine sur lequel ils communiquent pour le marché. De là à penser que tous les assurés aient été aussi largement récompensés... Voyons maintenant comment Solvabilité II traite le risque de rachat.

Avant de parler du risque de rachat en tant que tel, évalué pour les besoins en solvabilité, rappelons que Solvabilité II réforme également les calculs de provisions. Dans un bilan au format économique, les provisions techniques seront estimées avec des hypothèses de type *best estimate*, quand elles sont aujourd'hui estimées avec des hypothèses prudentes et conservatrices définies par exemple par le Code des Assurances. Quelles sont les implications pour ce qui concerne les rachats ? Aujourd'hui, le taux de rachat estimé dans les provisions techniques d'épargne est de 100% à chaque instant. L'entreprise d'assurances se doit donc d'immobiliser en permanence, dans ses comptes, le montant de l'épargne de chaque client, comme si le rachat devait intervenir demain. Dans un bilan plus économique, les provisions incluront des probabilités de rachat estimées en best estimate, c'est-à-dire reflétant au mieux la probabilité de rachat observée par l'assureur. Ainsi, les hypothèses de rachat qui ne servaient hier que dans le cadre des calculs d'embedded value ou d'études ALM (gestion actif-passif) vont-elles être introduites dans la comptabilité des entreprises d'assurances.

Ces lois de probabilités vont également servir à simuler le comportement des assurés dans le cadre des stress tests qui interviennent dans le calcul du SCR (Solvency Capital Requirement). Les assureurs sont encouragés à utiliser la meilleure connaissance possible qu'ils ont de leur portefeuille, pour modéliser les flux financiers de passif dans les scénarios de stress comme ceux des marchés financiers permettant d'estimer les SCR de taux ou d'actions. Les dernières pré-spécifications techniques de la cinquième étude quantitative d'impact (QIS 5) donnent des formules fermées (cf. TP 4.58) pour modéliser les rachats en fonction des garanties offertes, des taux servis, des conditions de marchés

financiers... Ces différentes formules doivent encore être calibrées par les assureurs pour refléter au mieux leur portefeuille.

Enfin, dans la liste des risques nécessitant la mise en oeuvre d'un calcul de SCR se trouve à part entière le risque de rachat à l'intérieur du module de risque de souscription. Dans les dernières parutions des pré-spécifications techniques, les assureurs sont priés d'étudier l'impact d'une hausse constante du taux de rachat de 50 % (limité à un taux de rachat de 100 %), l'impact d'une baisse constante du taux de rachat de 50 % (limité à un taux de rachat diminué de 20 %) et l'impact d'un rachat massif de 30 % (choc absolu) de la population sous risque. L'impact le plus significatif sera retenu pour être intégré au risque de souscription vie selon la matrice de corrélation définie dans les textes (paramètre de pseudo-corrélation égal à 50 % avec le SCR du risque de dérive des frais). Pour plus de détails sur toutes ces questions, l'organisme européen de contrôle EIOPA [*] (ex CEIOPS) est également une source intéressante de données.

∗. URL : https ://eiopa.europa.eu/

Chapitre 1

Etat de l'art

1.1 Intuitions sur les facteurs de risque potentiels

A première vue, il existe bien des facteurs pouvant affecter les comportements de rachat. Globalement ces facteurs de risque peuvent se résumer en deux grandes catégories, les effets structurels et les effets conjoncturels. Il est anecdotique de souligner que la liste des éléments influençant les rachats donnée ci-dessous est incomplète, puisque tout un chacun peut avoir ses propres motivations n'entrant pas dans ce cadre bien posé. De plus, les informations privées quant à l'anticipation des taux de rachat par les compagnies d'assurance ne sont généralement pas disponibles car il s'agit clairement d'une information stratégique. Cependant, nous pouvons dégager :

- parmi les facteurs *structurels* :
  - la **ligne d'affaire** et le **type de contrat** correspondant : épargne avec pure épargne, contrat mixte, unités de compte ; et prévoyance avec santé, maladies redoutées, incapacité-invalidité, décès ;
  - les **caractéristiques des contrats** : la participation aux bénéfices ; l'ancienneté du contrat ; le montant, le nivellement et la fréquence de la prime, le commissionnement, le réseau de distribution ;
  - **les caractéristiques des assurés** : le sexe, la catégorie socio-professionnelle, le lieu de résidence, le rapport prime sur salaire ou plus globalement la richesse de l'assuré, le statut fumeur, le statut marital, l'âge ;
- parmi les facteurs *conjoncturels* :
  - le changement de **législation** (voir annexe A.1) ;
  - le spread de taux de rendement avec la **concurrence** ;
  - **l'image** et le **rating** de la compagnie (voir annexe A.2) ;
  - l'évolution de l'offre (lancement de produits) et les **stratégies de vente** ;
  - les changements démographiques ;
  - l'évolution des taux d'intérêts, le taux de chômage, l'inflation, la croissance, le PIB. En fait cela représente tout ce qui est lié de près ou de loin au **contexte économique et financier**.

Plusieurs organismes ont réalisé des études empiriques sur les déclencheurs de rachat ; dont les principaux sont la Society Of Actuaries (SOA) aux Etats-Unis, la Fédération Française des Sociétés d'Assurance (FFSA) en France, et la Fellow Institute of Actuaries (FIA) au UK. Ces organismes publient des rapports (ex : FactBooks pour la SOA) qui ont généralement pour but d'apporter des réponses pragmatiques à des questions générales de type :

1. Quel est le niveau moyen de rachat **prévu** pour telle ligne de produit ?

2. Quel est le niveau moyen de rachat **constaté** sur cette même ligne de produit ?

3. Peut-on déterminer l'impact du niveau de prime sur les rachats ? Si oui, dans quelle proportion ?

Ces rapports sont précieux pour une bonne culture générale sur le sujet des rachats et permettent de confirmer certaines intuitions que les professionnels ont. Parmi ces résultats, nous pouvons citer quelques exemples extraits d'études réalisées par la SOA aux Etats-Unis afin d'en donner un aperçu : les taux de rachat des fumeurs seraient de 1,5 à 2 fois supérieurs à ceux des non-fumeurs (toutes choses égales par ailleurs), les personnes âgées rachèteraient moins que les jeunes... C'est également un moyen de connaître certaines pratiques courantes des compagnies d'assurance en matière de segmentation du risque de rachat ; citons pour cela un exemple disponible depuis un rapport de la SOA sur des produits "Return On Premium (ROP) Term life" *. Le tableau 1.1 résume comment les compagnies d'assurance considèrent l'évolution moyenne des taux de rachat sur cette ligne de produit. Ce type d'étude investigue ensuite plus précisément l'expérience des assureurs et leur avis. Par exemple sur les contrats ROP avec une couverture de 10 ans, les assureurs ont constaté que :

- les taux de rachat dans la 10ème année de contrat augmentent avec l'âge,
- les taux sont plus grands en l'année (L+1) qu'en l'année L (dans 30% des cas),
- le taux de rachat augmente sensiblement avec la somme assurée,
- dans la plupart des cas, les taux de rachat en montant sont plus élevés que les taux de rachat en nombre,
- les taux de rachat pour les hommes sont un peu plus grands que ceux des femmes,
- plus la périodicité des primes augmente, plus les taux de rachat sont élevés.

Nous voyons que ces études de cas permettent d'avoir certaines premières idées de segmentation ; ici la périodicité de la prime, l'âge et le sexe par exemple. Toutefois il faut veiller à garder un certain recul par rapport à ces résultats pour diverses raisons, notamment le fait que les compagnies d'assurance participant à ces études ne sont pas forcément représentatives de l'industrie (ou du marché dans lequel nous sommes positionnés). Les données sont de plus souvent imprécises et relativement peu fiables car aggrégées pour des raisons de confidentialité.

∗. Contrat temporaire décès avec garantie pour l'assuré de récupérer ses primes investies si l'événement ne se produit pas avant l'échéance du contrat.

Durée du contrat	Age à la souscription	L	L+1	L+2	L+3	L+4	L+5
10	Tout	80%	29%	17%	15%	15%	15%
15	Tout	83%	29%	18%	16%	16%	16%
20	Tout	82%	27%	17%	14%	14%	14%
30	Tout	83%	29%	19%	15%	15%	15%

TABLE 1.1 – Résumé des hypothèses de rachat pour des anciennetés de contrat allant de 0 à 5 ans (L à L+5). "Tout" comprend les personnes ayant moins de 35 ans et plus de 55 ans également.

1.2 Bibliographie sur la modélisation du rachat

Dans le monde académique, la modélisation des comportements de rachat a suscité un vif intérêt il y a une vingtaine d'années, avant de connaître un ralentissement. Historiquement, deux approches ont été privilégiées : l'hypothèse de la nécessité urgente de ressource pour l'assuré (Outreville (1990)) et l'hypothèse du taux d'intérêt (Pesando (1974) et Cummins (1975)). La première s'interprète facilement : admettons qu'un événement imprévu se produise dans la vie d'un assuré (achat d'une nouvelle voiture suite à un accident, achat d'un bien immobilier), celui-ci a besoin d'argent et va résilier son contrat d'assurance-vie afin de disposer des fonds nécessaires. L'hypothèse du taux d'intérêt est complètement différente : elle part du principe que si les taux d'intérêts du marché augmentent alors les taux de résiliation augmentent aussi car des opportunités d'arbitrage apparaissent naturellement sur le marché. Ainsi, des contrats à niveau de prime et de garantie égales offrent de meilleurs rendements.

Renshaw & Haberman (1986) sont les premiers à s'intéresser à la modélisation du comportement des assurés de manière statistique : ils analysent les comportements de rachat d'Assurance Vie en Ecosse en 1976 et dégage quatre principaux facteurs de risque de rachat que sont la compagnie, le type de contrat, l'âge et l'ancienneté du contrat. Ils utilisent des modèles linéaires généralisés (GLM) avec des termes d'intéraction entre ces facteurs de risque afin de bien modéliser l'hétérogénéité du portefeuille et les effets de l'ancienneté du contrat. Kim (2005) tente d'utiliser la régression logistique (GLM dénommé aussi "Logit") afin d'expliquer les rachats individuels d'un portefeuille coréen, en considèrant diverses variables explicatives catégorielles ou continues telles que l'âge, le sexe ou même le taux de chômage. Cette approche constituera d'ailleurs la première modélisation étudiée dans ce mémoire au chapitre 3. Dans le même esprit, Cox & Lin (2006) utilise un modèle Tobit (similaire au Logit mais avec un lien différent) et insiste sur l'importance de l'ancienneté du contrat comme facteur explicatif du rachat dans le cadre des taux de rachat de rentes. Un an auparavant, Kagraoka (2005) applique une loi de Poisson au cas de rachats de contrats d'assurance dommages au Japon. Pour modéliser la surdispersion de ses données, il réalise ensuite la même étude en utilisant une loi binomiale négative avec comme variables d'entrée le sexe, l'âge, la saisonnalité, l'ancienneté du contrat et le statut de travail. Dans ce contexte d'Assurance non-Vie, l'excellent mémoire (à paraître) de Dutang (2011) modélise les rachats par une approche élasticité prix, qui s'explique par la simple raison du terme du contrat fixé usuellement à un an.

Plus généralement, des résultats pour la modélisation d'évènements rares peuvent être trouvés dans l'excellent papier de Atkins & Gallop (2007) ; qui présentent plusieurs applications de modèles de régression associés à des données de comptage dont la loi de Poisson, la loi Binomiale Négative, et leurs extensions où le surplus de masse en 0 est modélisé ("zero-inflated"). Atkins & Gallop (2007) montrent que ces modèles sont adaptés lorsque les réponses présentent des distributions fortement asymétriques, et permettent d'éviter le biais introduit par les méthodes de régression par moindres carrés utilisées avec ce type de données (hypothèse de normalité des observations déraisonnable). En 2007, Fauvel & Le Pévédic (2007) rédigent un mémoire sur les rachats dans lequel l'ensemble des notions clefs sont abordées, avec une approche économiste via la théorie de l'espérance d'utilité couplée à des méthodes de finance quantitative. Le défaut selon nous de cette approche est qu'elle est basée sur la rationalité des assurés, une hypothèse relativement discutable.

Dans la littérature, d'autres auteurs comme Engle & Granger (1987) utilisent un

critère de minimisation des erreurs par la théorie des moindres carrés ordinaires lors du calibrage d'un modèle cointégré. Ils essaient de modéliser les rachats par des méthodes basées sur la notion de cointégration entre les taux de rachat et certaines variables économiques, dans le but de séparer la dynamique court-terme (de ces taux) des relations long-terme potentielles avec le taux de chômage et les taux d'intérêts. Tsai et al. (2002) démontrent une relation d'équilibre long-terme entre le taux de rachat et le taux d'intérêt, leur but final étant d'estimer le provisionnement nécessaire en tenant compte d'un taux de mortalité stochastique, de taux d'intérêts et de rachats précoces. Ils étudient en plus les questions de gestion de risque et de solvabilité de l'assureur. Dans un autre registre, Albert et al. (1999) étudient la relation entre les taux de rachat et les taux de mortalité entre 1991 et 1992 aux Etats-Unis avec des données provenant de multiples compagnies d'assurance comme Allstate, Equitable, Federal Kemper, John Hancock, Liberty Life Assurance, Minnesota Mutual, New York Life, Primerica, Sun Life et d'autres. Les données sont différenciées par statut fumeur.

Enfin une longue série d'auteurs s'intéresse ensuite à l'évaluation financière de l'option de rachat contenue dans les contrats d'Assurance Vie. C'est clairement le domaine dans lequel la littérature est la plus abondante, avec notamment l'école italienne. Pour n'en citer que quelques uns, Bacinello (2005) propose un modèle basé sur l'approche de Cox-Ross-Rubinstein (modèle CRR) et ses arbres binomiaux pour calculer la valeur de rachat de contrats à prime unique ou annuelle, avec une garantie plancher à maturité ou en cas de décès. Costabile et al. (2008) calculent le montant de primes périodiques associées à des contrats indexés sur les marchés financiers avec option de rachat et intérêts garantis, par le modèle CRR et un artifice de calcul (schéma avant-arrière couplé à une interpolation linéaire). Bacinello et al. (2008) se focalisent sur les rachats précoces et les considèrent comme des options américaines qu'ils valorisent grâce à un algorithme des moindres carrés Monte Carlo, à cause du fait que ces diverses options changent l'allure classique du payoff d'une option. Leur modèle prend en compte une mortalité stochastique et des sauts pour les indices financiers, le point fort de cet article est mis sur la performance de l'algorithme proposé. De Giovanni (2007) investigue plus particulièrement les différences dans la modélisation des comportements de rachat entre son approche, "l'espérance rationnelle" (Rational Expectation notée RE), et l'approche communément utilisée sur la place : la théorie American Contingent Claim (ACC) basée sur le comportement optimal des assurés (d'un point de vue de l'exercice de leur option). Il s'appuie sur le fait bien connu que les agents ne sont ni rationnels ni optimaux, et donne des résultats sur la différence d'impact des taux d'intérêts et de l'élasticité prix en utilisant l'approximation quadratique d'une fonction modélisant le comportement. Kuen (2005) utilisent l'approche ACC et le mouvement brownien pour décomposer les contrats avec participation au bénéfice en trois différentes options sous-jacentes : une obligation, une option de rachat et une option bonus. Il valorise ces contrats et arrive à la conclusion que la valeur de ce type de contrat est fortement sensible à l'option de bonus. Shen & Xu (2005) quantifient l'impact de l'option de rachat anticipé sur la valorisation de contrats en unités de compte (UC) à taux garanti, en utilisant le mouvement brownien géométrique et des équations différentielles partielles. Vandaele & Vanmaele (2008) explicitent la stratégie de couverture d'un portefeuille de contrats en UC comprenant une option de rachat en incorporant des hypothèses intéressantes : les temps de paiement et le temps de rachat ne sont clairement pas indépendants du marché financier, et un processus de Lévy modélise ce marché. Toutefois cet article n'est pas facilement abordable à cause de sa technicité. Récemment, Nordahl (2008) a écrit

un article intéressant sur la valorisation de l'option de rachat pour des contrats retraite. Il se base sur l'approche Longstaff-Schwartz et sur des simulations de Monte Carlo, en considérant le fait que l'option de rachat est comparable à une option américaine avec un strike stochastique. Torsten (2009) se place lui dans le cadre de contrats avec participation aux bénéfices de la compagnie, et souligne le fait qu'en général le problème de couverture et celui de valorisation ne peuvent être séparés. En effet le portefeuille d'investissement de l'assureur sert souvent de sous-jacent à la couverture, ce qui amène des difficultés supplémentaires.

La plupart des articles cités concerne la valorisation de ladite option de rachat, mais rares sont ceux dont l'objectif est la modélisation du comportement de rachat (notre intérêt). Par conséquent nous ne développerons pas de méthodes financières dans notre étude, mais c'est l'occasion de se rendre compte du clivage entre les deux principales écoles qui étudient les rachats : la première, composée essentiellement de chercheurs académiques (en nombre), s'attache à l'étude des aspects conjoncturels tandis que la seconde, menée par des professionnels (en nombre plus limité), se focalise davantage sur les aspects structurels. Pourtant ces deux aspects vont de pair et semblent aussi importants l'un que l'autre. La revue bibliographique manquant par définition d'exhaustivité, une très vaste littérature à considérer dans le cadre des rachats conjoncturels est d'ailleurs accessible via la modélisation des rachats de crédits en finance (Stanton (1995) ou encore Hin & Huiyong (2006)). A ce propos, Hin & Huiyong (2006) abordent un nouvel aspect de modélisation pour les comportements de rachat de crédits : l'utilisation de modèle de survie. Ils étudient mensuellement les rachats à Shangai entre 1999 et 2003 par l'usage du modèle à hasards proportionnels (de type modèle de Cox, Cox (1972)) dans le but de comprendre le fonctionnement du marché résidentiel chinois, avec des variables telles que le revenu des emprunteurs, le PIB. Ce papier détermine des facteurs de risque tout en développant l'aspect catégorisation des données d'entrée, mais ne traite malheureusement pas de la question des prévisions. Nous avons à titre personnel trouvé cette approche très intéressante, ce qui nous a poussé à l'implémenter. Néanmoins, nous ne présentons pas les résultats de ces études dans ce mémoire car la modélisation par analyse de survie ne nous a pas permis pour le moment de résoudre les problèmes rencontrés en pratique.

1.3 La gestion du risque de rachat chez AXA

Ce mémoire traite de la modélisation des **comportements** de rachat, bien que beaucoup de professionnels abordent le problème des rachats sous l'angle des montants rachetés plutôt que des décisions d'assurés. Ce regard est complètement légitime dans la mesure où le rachat d'un assuré très riche a évidemment plus d'impact que celui d'un assuré qui ne retirerait que peu d'argent. En termes de modélisation, les outils diffèrent suivant le phénomène étudié (lois de probabilité continues dans le cas des montants) mais les principaux facteurs d'influence sont identiques, d'où une transposition possible de notre travail à cette vision alternative. Nous décrivons dans cette section comment AXA gère le risque de rachat dans la majeure partie de ses entités ; en traitant successivement de l'opinion des équipes de terrain, de son ambition et des modélisations existantes.

1.3.1 Le point de vue des équipes Marketing

Nous avons trouvé pertinent de discuter avec les équipes du Marketing afin de connaître leurs attentes sur les débouchés que pourrait donner une bonne segmentation

et une bonne modélisation du risque de rachat. C'est également l'occasion de recueillir les intuitions des hommes de terrain sur le sujet. Nous remercions Xavier Blanchard de s'être prêté à ce petit jeu de questions-réponses et de nous avoir ainsi éclairé sur les points primordiaux dont il faut absolument tenir compte.

Nous définissons en premier lieu les grandes catégories de produit et l'importance associée aux comportements de rachat. Le tableau 1.2 se résume en deux idées simples : l'importance du risque de rachat diffère suivant les lignes d'affaire et il faut distinguer deux types de contrats en épargne. Les contrats avec garantie de taux de type fonds euros ou garantie plancher, et les contrats sans garantie de rendement avec les UC classiques par exemple. Les observations faites par les équipes Marketing tendent à montrer qu'une crise provoque globalement une baisse du nombre de rachats pour les contrats de la première catégorie (taux garantis), et une hausse du nombre de rachats pour les produits sans garantie de taux (la perception du risque de l'assuré change et il va réallouer les nouveaux flux d'épargne vers des supports garantis). La question du rachat est donc centrale pour tous les produits d'épargne individuelle ; mais ne l'est pas vraiment dans le domaine de la prévoyance où les contrats sont en majorité (environ 70 %) collectifs, or l'assuré ne connaît pas vraiment ses droits dans ce cadre-là et ne se pose généralement pas la question du rachat.

Trois principales dimensions expliquent les rachats selon les équipes de vente : la fiscalité, les pénalités de rachat et le réseau de distribution. L'effet de la fiscalité n'est évidemment visible que dans les pays pour lesquels certaines contraintes existent (ex : la France), les pénalités de rachat agissent sur l'assuré de la même manière qu'une contrainte fiscale (l'ancienneté du contrat guide souvent le profil de ces pénalités), et le commissionnement des agents de vente est une variable fortement liée au réseau de distribution. Un agent agréé sera tenté de provoquer le rachat aussitôt que les pénalités de rachat ou la fiscalité seront favorables à l'assuré, ou dès que le commissionnement qu'il reçoit pour la souscription de nouvelles affaires sur de nouveaux produits lui est favorable. Cette dernière remarque constitue d'ailleurs un véritable écueil en termes de modélisation, car il est impossible de prévoir à long terme la sortie de nouveaux produits et le comportement des agents de vente (bien que l'impact en soit majeur). Il est également important de garder en tête que le profil de rendement du produit va pousser l'assureur à favoriser ou non les rachats à un certain moment du contrat.

Rappelons que le rachat est un problème aussi bien à la hausse qu'à la baisse. Au Japon par exemple, AXA a dû transformer ses produits suite à l'explosion des taux de rachat dûe à la crise des taux d'intérêts (anormalement bas) à la fin des années 1990. Le niveau de la baisse ou de la hausse joue de manière relative car l'assuré réagit relativement à ce qu'il possède. Les équipes d'AXA s'attendent à un changement de

Lignes d'affaire	Catégorie de produit	Option de rachat	Importance des rachats
Epargne (Savings)	Pure épargne (Pure savings)	oui	très important
	Mixtes (Endowments)	oui	important
	Unités de compte (Unit-Linked)	oui	très important
	Index-Linked (IL)	oui	très important
Prévoyance (Protection)	Maladies redoutées (Critical Illness)	non	
	Dépendance (Long Term Care)	oui	peu important
	Garanties décès (Death benefit)	oui	peu important
	Rentes viagères (Whole Life)	oui et non (éviter l'anti-sélection)	peu important

TABLE 1.2 – Importance de la problématique des rachats pour les grandes catégories de produit.

comportement des assurés suite à la crise financière actuelle mais ce phénomène plutôt général reste difficile à quantifier car aucune donnée statistique n'est encore publiée. Une solution consiste à se procurer des études de marché réalisées en interne pour étudier la sensibilité des assurés aux taux d'intérêts (AXA France en a réalisé une récemment). Il y a communément en Marketing d'assurance une première segmentation en trois grandes classes selon la richesse des assurés : les "retail" (mass market), les "mass affluent" de richesse moyenne et les "affluent HNI". Cette dernière catégorie regroupe les personnes dont le rachat est le plus dangereux pour l'assureur, car la valeur de rachat est souvent très élevée et le remboursement de la provision mathématique du contrat se doit d'être immédiate.

Pour éviter les rachats massifs, certains produits (pour le moment marginaux) ont des clauses très spéciales comme un "facteur d'ajustement de marché" en cas de forte hausse des taux. Dans un futur proche, ce type de produit pourrait se développer dans la mesure où une des grandes peurs des assureurs aujourd'hui concerne la remontée soudaine des taux au sortir de la crise, qui pourrait entraîner une vague massive de rachats (il n'y a jamais eu de taux si bas que les taux actuels en Europe, d'où des nouvelles affaires souscrites depuis trois ans à de faibles rendements). En prévoyance, il est bien sûr très difficile de détecter les phénomènes d'antisélection et d'aléa moral hormis par la mise en place coûteuse de questionnaires et examens médicaux. D'un point de vue réputation, nous n'avons pas d'indicateur interne permettant de mesurer l'image de la compagnie et d'une manière générale, les équipes sont incapables de quantifier l'impact d'un scénario extrême, par définition non estimable. Selon le Marketing, notre étude s'inscrit donc surtout dans un cadre Solvabilité II, elle permettrait l'usage d'un modèle interne plus approprié qui permettrait éventuellement de dégager des bénéfices par une meilleure gestion actif-passif et de meilleures prévisions des rachats, grâce à l'évaluation en *best estimate* des provisions dédiées aux rachats.

1.3.2 Modèle interne : le groupe de travail sur les rachats

L'entité de gestion du risque du Groupe AXA (GRM) et l'ensemble des entités AXA de par le monde ont coorganisé en 2010 un groupe de travail afin de rassembler les expériences individuelles pour définir de nouvelles bonnes pratiques de gestion du risque de rachat. Le but était de développer un modèle interne simplifié avec des chocs suite à certaines recommendations de l'autorité de contrôle prudentielle (ACP) qui pointait globalement trois points à améliorer : une trop grande confiance en les avis d'expert, un besoin accru de calibration locale et la capacité à définir une distribution de probabilité pour les rachats.
En 2010, le risque de rachat arrivait en troisième position (d'un point de vue exposition) chez AXA derrière les risques liés aux dépenses et à la longévité, d'où la pertinence d'affiner sa modélisation. Le groupe de travail s'est concentré chronologiquement sur les questions suivantes : i) l'expérience passée des entités, ii) des discussions entre experts sur les chocs à appliquer (absolus ou relatifs, additifs ou multiplicatifs), iii) la calibration de ces chocs, et iv) la gestion (monitoring, reporting) de cette modélisation des rachats.
Les principales sources de **rachats massifs** identifiées par les entités sont :

- la peur de la faillite (image de la compagnie) : impact fort à moyen terme sur toutes les lignes d'affaire (ex : Japon fin des années 1990) ;
- les crises du marché financier : impact fort à court terme sur les contrats d'épargne (principalement les contrats en UC) ;
- un changement de législation : fort impact à court terme pour tout le monde ;

– l'évolution de l'environnement compétitif : impact fort à relativement long terme sur toutes les lignes d'affaire (ex : USA) ;
– un problème avec le réseau de distribution : fort impact à l'échelle du produit.

Il est très difficile de savoir ce que les assurés comptent faire avec l'argent récupéré lors du rachat puisqu'il s'agit d'une information privée, mais nous notons que la totalité de ces points (excepté le dernier) s'inscrivent dans des tendances conjoncturelles. En ce qui concerne la prise en compte des marchés financiers, les participants évoquent l'utilisation d'une fonction dynamique de modélisation des rachats (voir section 1.3.3). Le problème de la corrélation entre produits est également abordé mais cette question reste ouverte pour le moment du fait de sa complexité, laissant libre choix quant à la granularité des études.

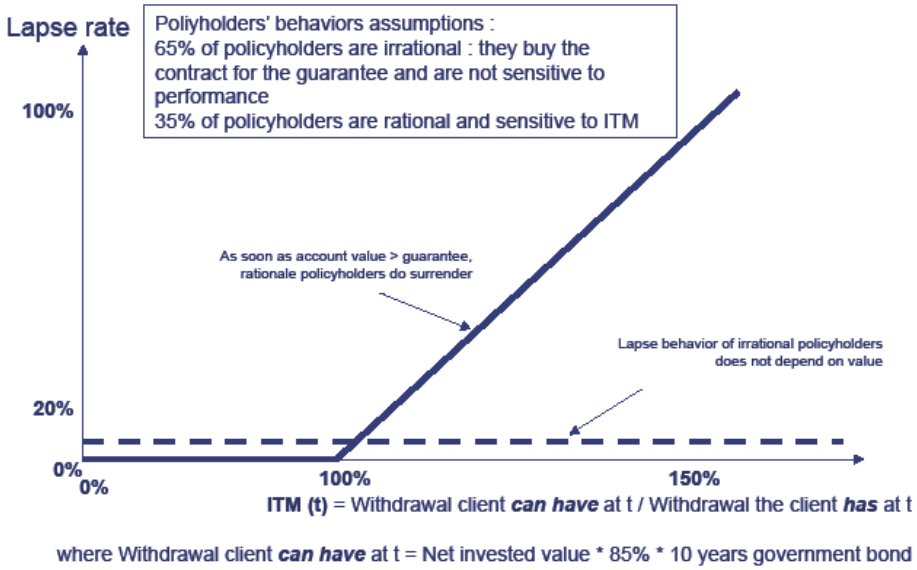
Aujourd'hui, le Groupe a identifié un ensemble de facteurs qui guident les décisions de rachat, mais n'utilise pas de modélisation à proprement parler. Pour définir le risque de base du rachat, les entités se basent sur les données empiriques segmentées par des facteurs comme l'ancienneté, le réseau de distribution, le type de prime, le profil des pénalités de rachat et le segment de la clientèle (ce qui donne des tableaux bidimensionnels sachant que chaque entité ne prend en général en compte que deux de ces facteurs, comme vu dans le tableau 1.1). Dans le domaine des assurances collectives, l'usage commun ne préconise aucune segmentation. Les hypothèses de rachat pour le business individuel sont spécifiées pour des groupes de produits similaires, exceptionnellement pour un produit en particulier s'il est destiné à un très large public. Il est très difficile d'obtenir les données de rachat par contrat (ou par tête), ce qui simplifie l'analyse mais ne permet pas une granularité suffisante pour parvenir à segmenter les assurés (heureusement nous avons pu récupérer des bases de données par contrat pour quatre entités dans notre étude). L'idée qui ressort des discussions d'experts du rachat au niveau du Groupe est qu'il faut définir un scénario de masse (modélisation dans le cadre général, risque de base), auquel nous greffons des scénarios de chocs qui interviennent dans un cadre exceptionnel (risque occasionnel), tout en limitant la complexité du modèle. Cette idée de chocs très importants tels que définis dans le QIS 5 doit toutefois être modérée : le constat d'AXA lors de la crise financière est en accord avec une augmentation marquée de la volatilité des taux de rachat, mais pour la plupart des produits l'amplitude des chocs n'est pas aussi grande que présumée. Nous nous intéressons dans ce mémoire à la calibration de ces deux types de risque (de base et occasionnel) qui débouchera sur une modélisation probabiliste du taux de rachat, donnant ainsi accès à une distribution de probabilité.

1.3.3 Fonctions dynamiques d'ajustement

La gestion du risque de rachat dans le groupe AXA repose sur une méthode relativement simple : tableau à double (voire multiples) entrée(s) pour le risque de base, auquel nous appliquons une fonction de rachat dynamique qui permet d'ajuster ce risque de base en fonction des conditions de marché (concurrentiel, financier). Typiquement, le risque de base est défini comme dans le tableau 1.1. En fonction des entités, la fonction dynamique prend usuellement une des quatre formes présentées ci-dessous, dont les deux premières sont basées sur la valorisation de l'option de rachat.

Fonction linéaire Cette fonction dynamique se base sur l'évaluation financière de l'option de rachat et ajuste le taux de rachat de base en fonction de la valeur de cette option. Pour cela, le but est de déterminer si l'option est dans la monnaie ou non :

FIGURE 1.1 – Exemple de fonction dynamique des rachats de forme linéaire.



dès que l'option est dans la monnaie, on augmente linéairement (figure 1.1) le taux de rachat pour la part des assurés qui est estimée sensible au marché.

Fonction en escalier Lorsque la valeur de l'option de rachat est élevée, l'assuré est supposé avoir tout intérêt à la conserver et le niveau de rachat est bas. Le raisonnement inverse est valable et suppose la définition d'un facteur multiplicatif (ici 10 dans la figure 1.2) qui provoque une forme en escalier et donc un changement brutal du taux.

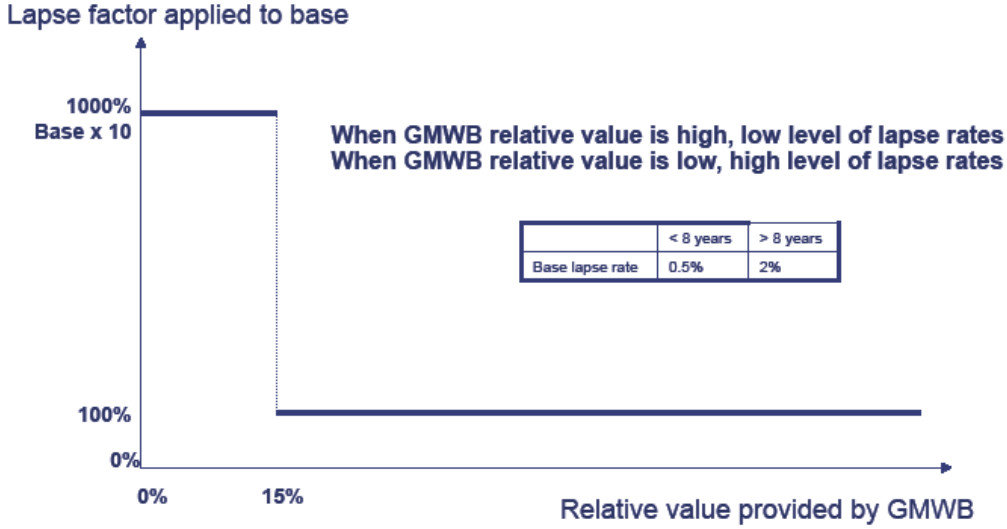
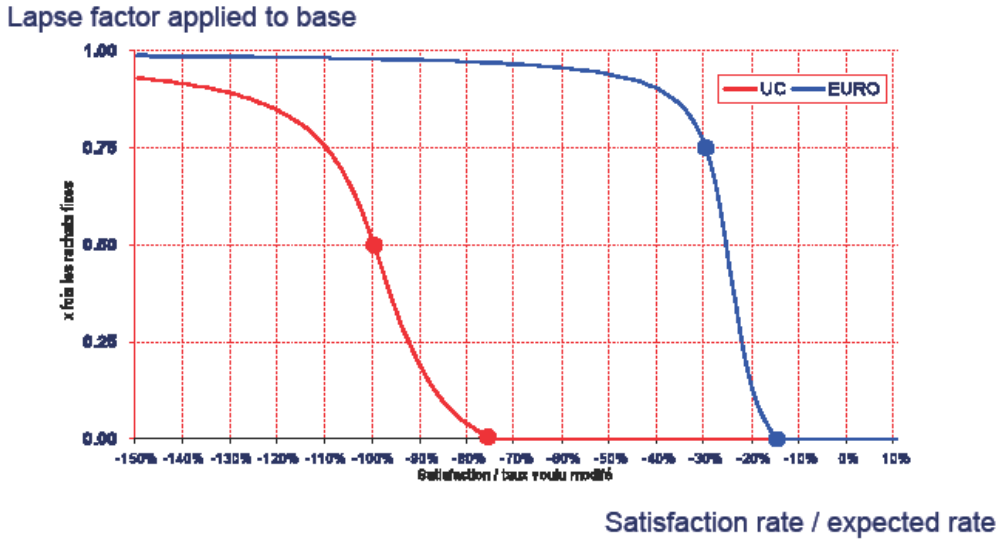


FIGURE 1.2 – Exemple de fonction dynamique des rachats en escalier.

Fonction exponentielle La fonction exponentielle qui s'applique au risque de base est construite suivant la valeur du *spread* (rendement espéré - taux crédité) :

FIGURE 1.3 – Exemple de fonction dynamique des rachats de forme Arctangente.



– si ce *spread* est négatif, alors nous considérons la fonction *exp(slope* × *spread)*, où *slope* est en fait la vitesse de convergence vers le rendement espéré de l'assuré (par défaut 50%)
– si ce *spread* est positif, alors nous considérons la fonction *2 - exp(-slope* × *spread)*.

Si l'on devait se représenter ces fonctions, nous verrions que c'est toujours la même idée sous-jacente : si le contrat est favorable à l'assuré, le taux de rachat est abaissé par rapport à son niveau de base et inversement.

Fonction arctangente L'idée ici rejoint la modélisation dynamique linéaire ou en escalier, mais introduit une subtilité quant à la sensibilité des assurés par rapport à la différence entre leurs attentes et ce qu'ils reçoivent. Cette sensibilité non linéaire (figure 1.3) est caractérisée par un seuil à partir duquel le facteur multiplicatif "explose". C'est la modélisation dynamique la plus courante, et c'est celle dont nous discuterons les hypothèses au chapitre 3 (courbe en S).

1.4 Conclusion

Nous pouvons dégager de la revue bibliographique quatre grands types de modélisation : une modélisation financière sous forme de valorisation d'option (modélisation individuelle de la décision), une modélisation statistique sous forme de série temporelle (modélisation collective des décisions de rachat), une modélisation économétrique (individuelle) basée sur la théorie de l'espérance d'utilité et une modélisation probabiliste (individuelle) sous forme de modèle GLM. Nous privilégierons cette dernière approche car elle permet notamment de prendre en compte les caractéristiques individuelles et de modéliser la **décision** de rachat sans hypothèse préalable sur la rationalité des comportements. Nous terminons cette introduction par la définition claire des objectifs de ce mémoire : nous nous orienterons sur la modélisation des comportements de rachats suivant le type de contrat, la localisation géographique et toute caractéristique identifiable qui pourrait avoir un impact probant sur les comportements de rachat. Les recherches bibliographiques de même que l'implémentation et l'amélioration de modèles

opérationnels existants nous aideront à déboucher ensuite sur de nouveaux choix de modélisation.

Dans le chapitre 2, nous abordons la problématique de segmentation du risque de rachat par l'usage de deux modèles complémentaires. Ce sera l'occasion de mettre en exergue dans une étude de cas certains facteurs de segmentation importants et de créer des classes de risque. En chapitre 3, nous discutons de l'hypothèse d'indépendance des comportements en illustrant ses conséquences sur un exemple concret. Nous en proposons ensuite une alternative qui permet de mieux refléter la réalité par l'introduction d'un modèle à chocs communs. Certains résultats qualitatifs intéressants y sont démontrés. Le chapitre 4 lie les idées développées en chapitres 2 et 3 : nous présentons la théorie des modèles mélange appliqués à notre problématique, et proposons une méthodologie d'étude pour la modélisation probabiliste des comportements de rachat sur un portfeuille réel d'Assurance-Vie. Enfin le chapitre 5 traite de la validation de cette méthodologie par des applications multiples et variées sur le portefeuille espagnol d'AXA, agrémentées de la description succinte des produits étudiés et de quelques statistiques descriptives en guise de support à la modélisation finale.

Chapitre 2

Segmentation du risque de rachat

Comme nous l'avons déjà évoqué, la compréhension de la dynamique des taux de rachat est cruciale pour les compagnies d'assurance qui doivent faire face à plusieurs problèmes qui y sont liés. Il y a tout d'abord la problématique des rachats anticipés qui entraînent l'impossibilité pour la compagnie de recouvrir ses frais d'émission, de gestion et d'administration du nouveau contrat (environ 3,5%). En effet, l'assureur paie ces frais avant ou à l'émission du contrat et espère faire des profits au cours de la vie du contrat, ces profits n'étant pas réalisés en cas de rachat précoce. Nous constatons ainsi que le profil temporel du rachat a une importance toute particulière, puisque de ce profil vont dépendre les coûts du rachat pour l'assureur. De plus, les assurés qui ont certains problèmes de santé et d'assurabilité auront tendance à ne pas racheter leur contrat, causant finalement plus de sinistres que prévu (phénomène d'anti-sélection). Enfin il existe toujours le risque de taux d'intérêts : au cours de la vie des contrats, ces taux varient. Plaçons nous par exemple dans un contexte de contrat d'épargne à taux garanti : si les taux d'intérêts s'effondrent, l'assureur doit tenir ses engagements et verser aux assurés un taux garanti supérieur au rendement de ses actifs, le risque étant donc que les rachats soient beaucoup moins nombreux que prévu et que l'assureur manque de liquidité. Inversement, les assurés seront plus à même de racheter leur contrat en cas de hausse des taux car les nouveaux contrats offriront de meilleurs rendements à niveau de garantie équivalent. L'assureur devra donc rembourser aux assurés la valeur de rachat de ces contrats dans un contexte où l'emprunt d'argent peut s'avérer très coûteux ! Finalement l'assureur peut subir une série d'effets indésirables en cascade : pas le temps de recouvrir ses frais, obligation d'emprunter à prix fort et nécessité de liquider ses actifs au pire moment (cependant les rachats ne sont pas qu'une mauvaise nouvelle pour l'assureur puisque celui-ci se débarrasse de garanties qui ont un coût). Heureusement les observateurs noteront que l'agent (assuré) n'est en général ni rationel ni optimal, même si ces comportements sous-optimaux tendent à disparaître du fait d'une information toujours plus accessible.

Toutes ces considérations montrent que l'enjeu d'une modélisation précise des comportements de rachat est primordiale en termes de profitabilité et de solvabilité pour l'assureur. Les praticiens fixent des hypothèses de rachat, fruit de l'étude statistique de la collection de données expérimentales rendue complexe par l'essence même de celles-ci pour plusieurs raisons : les types de données, leur dimension, la gestion des données manquantes...Le défi est donc de sélectionner le minimum de données apportant le maximum d'information, ce que nous essayons de faire dans ce chapitre par l'usage de deux modèles complémentaires de segmentation : les arbres de classification et de régression (CART) et la régression logistique (LR).

La méthode CART développée par Breiman et al. (1984) et la LR (Hilbe (2009)) nous ont permis de prouver avec différents portefeuilles d'Assurance-Vie d'AXA le pouvoir discriminant de certaines caractéristiques sur la décision de rachat. Nous présentons rapidement dans un premier temps les fondamentaux de chacun des modèles ainsi que leurs hypothèses et limites. Nous discutons au final des différences entre les deux modélisations en termes de résultats numériques et d'un point de vue opérationnel, et justifions l'emploi d'autres modélisations plus ajustées dans la suite du mémoire. Le but de ce chapitre est donc de i) réduire la dimension de l'espace des variables à prendre en compte dans la future modélisation, ii) déterminer quelle méthode semble la plus adaptée en regardant les taux d'erreur de classification, iii) trouver les déclencheurs essentiels du rachat en régime de croisière (économique).

Nous gardons à l'esprit que cette segmentation ne représente pas la réalité en période de crise (financière, d'image) et introduit un biais non-négligeable car nous n'y considérons pas le contexte économique. Les effets de cohorte ne sont également pas pris en compte puisque la date de rachat n'est pas introduite en facteur explicatif. Nous reviendrons sur ces remarques pour proposer des extensions possibles dans la suite du mémoire lors de prévisions futures de taux de rachat incluant des facteurs dynamiques. Cette première segmentation est utile à plusieurs titres : elle permet de mieux comprendre les comportements assurés, de planifier une segmentation du risque de rachat et d'améliorer le design de nouveaux produits (hypothèse de taux moyen de rachat, clauses et options des contrats). L'exemple considéré sert ici à alimenter la théorie pour une meilleure compréhension, les résultats pour l'ensemble des produits sont donnés dans le chapitre 5.

2.1 Modélisation CART

La méthode CART, outil non paramétrique flexible, est basée sur un algorithme à la fois itératif et récursif. Développée par Breiman et al. (1984) dans le but de diviser les données d'origine à l'aide de règles déterministes, ses arbres binaires offrent une manière puissante et conviviale de fournir des résultats dans les problèmes de classification. La particularité de l'algorithme CART en comparaison à ses "congénéres" est qu'il n'existe pas de règle d'arrêt lors de la construction de l'arbre. De manière générale, les deux principaux buts d'un processus de classification sont de produire un bon classifieur et d'avoir un bon pouvoir prédictif. Nous entendons par "bon" une procédure qui amène à des erreurs acceptables, bien que nous verrons qu'un arbitrage entre ces deux notions doit être fait par l'utilisateur dans le sens où un gain en précision de classification entraîne généralement une perte de pouvoir prédictif. CART se révèle être très utile, mais l'utilisateur doit être conscient que plusieurs modèles de segmentation doivent être utilisés dans l'idéal pour obtenir un résultat robuste.

2.1.1 Le modèle

Nous présentons dans cette section comment construire l'arbre. La figure B.1 des annexes B.1 indique et détaille les différentes étapes à suivre, ainsi que les concepts sous-jacents pour le lecteur que cela intéresserait. Nous trouvons intéressant de fournir une chronologie claire de l'algorithme CART car elle n'apparaît pas explicitement dans la littérature.

Construction de l'arbre de classification

Notation Soit $\epsilon = (x_n, j_n)_{1 \leq n \leq N}$ un échantillon de taille N, où les j_n représentent les observations de la variable réponse Y (Y $\in C = \{1, 2, ..., J\}$ et les $x_n = \{x_{n_1}, x_{n_2}, ..., x_{n_d}\}$ sont les observations de X dans $\mathbb{X}$, ensemble des d variables explicatives ($\mathbb{X} = \prod_{i=1}^{d} \mathbb{X}_i$ où $\mathbb{X}_i$ est un ensemble de variables continues et/ou catégorielles). Soit

- $\forall x \in \mathbb{X}$, le classifieur $class(., \epsilon)$ classe x dans un groupe j $\in$ C.
- Le groupe j à-priori est défini par $\pi_j = \frac{N_j}{N}$ où $N_j = card\{j_n | j_n = j\}$.
- Sachant $t \subset \mathbb{X}$ (t sous-ensemble fini de $\mathbb{X}$), notons N(t) = card$\{(x_n, j_n) \in \epsilon, x_n \in t\}$.
- $N_j(t)$=$card\{(x_n, j_n) \in \epsilon, j_n$=j sachant que $x_n \in t\}$.
- Un estimateur par substitution de P(j,t), noté p(j,t), est donné par p(j,t) = $\pi_j \frac{N_j(t)}{N(t)}$.
- Un estimateur par substitution de P(t), noté p(t), est donné par p(t) = $\sum_{j=1}^{J} p(j,t)$.
- P(j | t), la probabilité à-posteriori d'appartenir à la classe j, est estimée par $\frac{p(j,t)}{p(t)} = \frac{N_j(t)}{N(t)} = \frac{p(j,t)}{\pi_j}$.

Comment débuter ? Le principe est de diviser $\mathbb{X}$ en q classes, où q n'est pas connu à l'avance (*a-priori*). La méthode construit une séquence croissante de partitions de $\mathbb{X}$; On passe d'une partition à l'autre en appliquant des *règles de division binaires* telles que :

$$x \in t, \text{ avec } t \subset \mathbb{X}.$$

Par exemple, la première partition de $\mathbb{X}$ peut être le sexe de l'assuré. L'assuré dont la caractéristique est x est soit une femme soit un homme (une spécification des règles binaires est détaillée en annexe B.1.3).

Nous commençons par diviser la *racine* $\mathbb{X}$ en deux sous-ensembles disjoints appelés *noeuds* et notés t_L et t_R. Chaque noeud est ensuite divisé de la même manière (s'il contient au moins deux éléments). Au final nous obtenons une partition de $\mathbb{X}$ en q groupes appelés *noeuds terminaux* ou *feuilles*.
Dans la suite, nous notons $\tilde{T}$ l'ensemble des *feuilles* de l'arbre T ; T^t est l'ensemble des *descendants* de *l'ancêtre* t dans l'arbre T (voir l' illustration en figure B.2).

Nous mesurons la qualité de la division d'un noeud t en t_L et t_R grâce à un *critère d'impureté*. Ce concept est également expliqué en détail en annexe B.1.3. Dans notre cas, l'impureté du noeud t dans l'arbre T est la quantité

$$impur(t) = g(p(1|t), p(2|t), ..., p(J|t)), \tag{2.1}$$

où g est la fonction d'impureté. Par conséquent, l'impureté de l'arbre T est donnée par

$$Impur(T) = \sum_{t \in \tilde{T}} Impur(t) \tag{2.2}$$

où $Impur(t) = p(t) impur(t)$.
Une règle de division Δ d'un noeud t donne p_L=$p(t_L)/p(t)$ observations dans t_L et p_R=$p(t_R)/p(t)$ observations dans t_R. Nous aimerions maximiser la *pureté* :

$$\delta\ impur(\Delta, t) = impur(t) - p_L impur(t_L) - p_R impur(t_R) \tag{2.3}$$

La pureté de l'arbre est censée augmenter à chaque division, ce qui impose la contrainte naturelle suivante :

$$impur(t) \geq p_L\ impur(t_L) + p_R\ impur(t_R).$$

Respectons nous toujours cette inégalité ? La réponse est "oui" si g est concave. Dans la plupart des applications (y compris la notre), nous considérons l'index de diversité de Gini, interprétable comme une probabilité de mauvaise classification. C'est la probabilité d'affecter la classe k à une observation choisie aléatoirement dans le noeud t, multipliée par la probabilité estimée que cette observation appartienne en réalité à la classe j. Il existe aussi d'autres fonctions d'impureté qui ont une interprétation encore plus simple (annexe B.1.3), mais il n'existe pas de justification particulière pour l'usage de telle ou telle fonction (en particulier elles sont toute concaves, et les propriétés de l'arbre final ne sont pas vraiment impactées par ce choix, comme décrit dans Breiman et al. (1984)). Traditionnellement la division optimale $\Delta_t^\star$ d'un noeud t satisfait

$$\Delta_t^\star = \underset{\Delta \in D}{argmax}\ (\delta\ impur(\Delta, t)), \tag{2.4}$$

où $argmax\ (\delta\ impur(\Delta, t))$ désigne la règle de division Δ qui maximise $\delta\ impur(\Delta, t)$.

Le processus génère donc une décroissance d'impureté aussi rapide que possible à chaque étape. Intuitivement, cela signifie qu'un maximum d'observations doivent appartenir à la même classe dans un noeud donné, ce qui définit la règle de division à choisir. Maximiser le gain de pureté (ou d'homogénéité) par la division du noeud t revient à maximiser le gain de pureté de l'arbre T. Nous obtenons ainsi un arbre T' (voir en annexe la figure B.2) plus ramifié en partant de l'ancêtre t vers les descendants (t_L,t_R) par Δ, et (2.2) donne

$$Impur(T') = \sum_{w \in \tilde{T} - \{t\}} Impur(w) + Impur(t_L) + Impur(t_R).$$

Ce qui donne une fluctuation d'impureté au niveau de l'arbre T de

$$\begin{aligned} F &= Impur(t) - Impur(t_L) - Impur(t_R) \\ &= \delta Impur(\Delta, t) \\ &= p(t)\delta impur(\Delta, t). \end{aligned} \tag{2.5}$$

Il s'agit donc de la probabilité d'être présent dans ce noeud multipliée par la décroissance d'impureté donnée par Δ. L'étape suivante consiste à définir quand arrêter les divisions, ce qui relève du choix de l'utilisateur. Certaines règles d'arrêt sont naturelles tandis que d'autres sont purement arbitraires : i) les divisions s'arrêtent évidemment lorsque les observations des variables explicatives dans une classe donnée sont identiques ; ii) définir un nombre minimal d'observations dans un noeud (plus ce nombre est petit et plus le nombre de feuilles sera grand) ; iii) choisir un seuil λ de décroissance minimum de l'impureté : soit $\lambda \in \mathbb{R}_+^*$,

$$\underset{\Delta \in D}{max}\ \delta\ Impur(\Delta, t) < \lambda \Rightarrow \text{arrêter la division}$$

En fait et comme énoncé au début de cette section, il n'y a pas de régle d'arrêt dans les CART ; on construit l'arbre le plus ramifié et on l'élague ensuite par une procédure avancée détaillée en annexe B.1.3.

La fonction de classification

Le but est de construire un classifieur, noté $class(., \epsilon)$, tel que

$$\begin{aligned} class\ &:\ \mathbb{X} \to C \\ &\qquad x \to class(x, \epsilon) = j, \end{aligned}$$

où $B_j = \{x \in \mathbb{X}; class(x, \epsilon) = j\}$, pour classer les assurés (sachant leurs caractéristiques "x") dans un ensemble B_j afin de prédire la réponse. Cette fonction doit si possible classer au mieux les données et avoir un pouvoir prédictif intéressant. Considérons que l'arbre optimal a été construit ; pour connaître la classe d'appartenance d'un noeud terminal, nous utilisons la règle

$$class(x, \epsilon) = \underset{j \in C}{argmax}\ p(j|t), \tag{2.6}$$

autrement dit la fameuse règle de *Bayes* qui maximise la probabilité *à-posteriori* d'être dans la classe j sachant que nous sommes dans le noeud t. Ce processus nous permet ainsi d'effectuer des prévisions de classification.

Une estimation de la mauvaise classification d'une observation dans le noeud t (par rapport à la classe observée) est donnée par

$$r(t) = 1 - class(x, \epsilon) = 1 - \underset{j \in C}{max}\ p(j|t), \tag{2.7}$$

Soit $\hat{\tau}(t) = p(t)r(t)$ le taux de mauvaise classification du noeud t. Pour chaque noeud, c'est la probabilité d'être dans le noeud t multipliée par la probabilité de mal classer une observation sachant que nous sommes dans ce noeud t. Nous en déduisons immédiatement le taux global de mauvaise classification de l'arbre T, donné par

$$\hat{\tau}(T) = \sum_{t \in \tilde{T}} \hat{\tau}(t). \tag{2.8}$$

Finalement, nous pouvons résumer les quatre étapes essentielles de la procédure de construction de l'arbre :

1. un ensemble de questions binaires $\{x \in S\,?\}$, $S \in \mathbb{X}$,
2. une fonction d'impureté pour le critère de qualité d'ajustement (choix arbitraire),
3. une règle d'arrêt des divisions (choix arbitraire),
4. une procédure de classification permettant d'affecter à chaque feuille une classe.

Comme vu précédemment CART construit un arbre maximal T_{max} et l'élague, ce qui permet d'éviter une règle d'arrêt arbitraire des divisions.

Estimation de l'erreur de prévision

L'*erreur de prévision* est évaluée par la probabilité qu'une observation soit mal classée par $class(., \epsilon)$, c'est-à-dire :

$$\tau(class) = P(class(X, \epsilon) \neq Y)$$

L'efficacité du prédicteur est basée sur l'estimation de cette erreur. Le taux de mauvaise classification réel $\tau^*(class)$ ne peut pas être estimé lorsque la procédure de classification est construite à partir de l'ensemble des données, mais il existe plusieurs estimateurs dans la littérature (Ghattas (1999)). L'expression du taux de mauvaise classification dépend de l'échantillon d'apprentissage (détails en annexe B.1.3) :

- estimation par **resubstitution** du taux de mauvaise classification de l'arbre : nous considérons toutes les observations ϵ pour l'échantillon d'apprentissage. Les résultats ne sont pas représentatifs car nous classons les mêmes données que celles qui ont servi à la construction du classifieur. C'est donc évidemment le pire estimateur.

- estimation par **échantillon témoin ou de validation** : soit W $\subset \epsilon$ l'échantillon témoin de taille $N' < N$ (N est la taille de ϵ et en général $N^{'}$=$N/3$). Nous construisons le classifieur avec l'échantillon d'apprentissage et validons son efficacité sur l'échantillon témoin. Cet estimateur est meilleur mais nécessite beaucoup de données.
- technique des **validations croisées** : supposons ϵ divisé en K sous-groupes disjoints $(\epsilon_k)_{1\leq k\leq K}$ de même taille. Définissons K jeux de données d'apprentissage tels que $\epsilon^k = \epsilon - \epsilon_k$. L'idée est de construire une procédure de classification sur chaque ϵ^k telle que $class^k(.) = class(., \epsilon^k)$. Cette technique est recommandée avec peu de données disponibles car le taux d'erreur (moyenne des K taux d'erreur) est plus réaliste.

Dans la suite $\tau(T)$ est l'erreur de prévision sur T ; $\hat{\tau}(T)$, $\hat{\tau}^{ts}(T)$ et $\hat{\tau}^{cv}(T)$ leurs estimations, respectivement aux trois méthodes ci-dessus.

2.1.2 Limites, améliorations

La classification par arbre offre des avantages certains : i) pas de restriction sur le type de données (catégorielles ou continues) ; ii) les résultats finaux sont simples à interpréter et à visualiser ; iii) l'algorithme induit une méthode pas-à-pas automatique de sélection de variable et donc une réduction de la dimension de l'espace et de sa complexité. De plus, les transformations monotones des variables ne changent pas les résultats, et l'aspect non-paramétrique ne suppose pas de relation prédéterminée entre la variable réponse et les variables explicatives. Les intéractions entre prédicteurs sont en général bien identifiées.

Cependant, quelques inconvénients subsistent parmi lesquels le fait que les divisions soient basées sur une seule variable alors que nous pourrions penser que des combinaisons de variables seraient parfois plus adéquates, auquel cas l'algorithme serait mauvais pour représenter la structure des données. Nous pouvons également citer le fait que l'effet d'une variable peut être caché par une autre lors du choix des règles de division. Il existe des solutions pour éviter ce phénomène, comme classer l'effet potentiel de chaque variable explicative lors de la division : ce sont les *secondary* et *surrogate splits* dans la littérature (également utilisées pour des données manquantes, voir Breiman et al. (1984)). L'arbre final peut aussi être difficile à utiliser en pratique car trop ramifié (cas extrême : une observation par feuille, ce qui amène un taux de mauvaise classification nul qui n'est évidemment pas du tout réaliste !), mais l'utilisateur peut jouer sur la taille de l'arbre par l'introduction d'un coût de complexité dans l'algorithme d'élagage (annexe B.1.3) pour pallier à ce problème.

Enfin, bien que CART donne une idée claire de l'importance de chaque variable explicative par lecture de l'arbre final depuis la racine jusqu'aux feuilles, Ghattas (2000*b*) critique un déficit de robustesse dans les résultats : une légère modification des données peut engendrer différents classifieurs, une contrainte importante dans la problématique de prévision. Nous voudrions évidemment éviter ce type de comportement pour lequel une variable peut s'avérer déterminante dans le processus de classification avec un certain jeu de données et être absente dans un jeu de données quasi-similaire. Les solutions développées à ce propos sont la validation croisée (Ghattas (2000*a*)), les *bagging predictors* ou *arcing classifiers*. Ces deux dernières techniques sont une aggrégation de type *bootstrap* de classifieurs construits sur des échantillons bootstrap, et leur robustesse et significativité ont été validées dans diverses études ((Breiman (1996), Breiman (1994) et Breiman (1998)). Elles ont amené au développement des “forêts aléatoires” (Breiman (2001)), un algorithme que nous utiliserons dans nos applications. Pour plus de détails,

consulter la page web de Breiman, la documentation de la librairie R `randomForest` * qui y est dédiée et Breiman et al. (1984).

2.2 Segmentation par modèle logistique (Logit)

La régression logistique (Hosmer & Lemeshow (2000), Balakrishnan (1991)) appartient à la classe des modèles linéaires généralisés (McCullagh & Nelder (1989)). Elle permet de modéliser la probabilité d'occurence d'un événement binaire à partir de données observées (les covariables) catégorielles ou continues, en ajustant une courbe logistique aux données. Ce modèle de choix est utilisé dans le cadre des régressions binomiales, principalement dans le domaine médical et dans le monde du marketing. Les actuaires l'utilise également pour modéliser la mortalité (qui présente une forme exponentielle en fonction de l'âge, pas loin de la forme logistique pour de petites probabilités) avec les données empiriques de leur portefeuille, dans le but de le segmenter. Dans notre contexte, l'objectif est de segmenter la population par rapport au risque de l'événement binaire représenté par la décision de rachat. La présentation théorique sera raccourcie étant donné la popularité de cette modélisation, quelques exemples d'application sont consultables dans Kagraoka (2005) ainsi que des modèles similaires dont le modèle Tobit (Cox & Lin (2006)) ou le modèle de Cox (Cox (1972)). Pour de plus amples comparaisons de ces différents modèles, l'article d'Austin (2007) est une référence intéressante.

2.2.1 Pourquoi utiliser la régression logistique ?

La fonction logistique est très utile car elle permet d'obtenir une image $\Phi(z)$ dans [0,1] à partir d'un antécédent z prenant des valeurs sur l'ensemble de la droite des réels :

$$\Phi(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}. \tag{2.9}$$

Notre volonté de modéliser une probabilité de rachat entre complètement dans ce cadre-là, sachant de plus que la propriété de non-décroissance d'une fonction de répartition classique est respectée par la fonction logistique. z représente l'exposition à un ensemble de facteurs de risque et est appelé *prédicteur linéaire*. Il est donné par l'équation de régression classique

$$z = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k,$$

où les X_i sont les covariables (explicatives), par exemple l'âge. Ainsi $\forall i = 1, ..., k; \beta_i$ représente le coefficient de régression associé au facteur de risque i. Nous noterons les coefficients de régression $\beta = (\beta_0, ..., \beta_k)^T$ et les variables $X = (X_1, ..., X_k)^T$.
Si l'on considère une approche stricte de régression, l'idée est de transformer la sortie d'une régression linéaire classique pour obtenir une probabilité en utilisant une fonction de lien (ici le "logit-link", mais il existe aussi le lien "probit").

2.2.2 Estimation des paramètres par maximum de vraisemblance

Nous avons vu que les rachats sont binomialement distribués (nombre de rachats $\sim B(n, \Phi(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k))$, où n est le nombre d'individus). La méthode du maximum de vraisemblance (ML) nous permet d'estimer de manière classique les

∗. Disponible à http ://cran.r-project.org/web/packages/randomForest/index.html

paramètres de la probabilité de rachat. Par définition, la fonction de vraisemblance pour une loi binomiale vaut

$$L(\beta, X) = \prod_{i=1}^{n} \Phi(X_i\beta')^{Y_i}(1 - \Phi(X_i\beta'))^{1-Y_i},$$

où Φ est définie dans (2.9). La log-vraisemblance est donc

$$\ln(L(\beta, X)) = \sum_{i=1}^{n} Y_i(X_i\beta') - \ln(1 + e^{X_i\beta'}), \tag{2.10}$$

L'estimateur du maximum de vraisemblance $\hat{\beta}$ satisfait $\dfrac{\partial \ln(L)}{\partial \hat{\beta}}$=0. Cette condition amène généralement à un système d'équations dont la solution n'est pas explicite et n'admet donc pas de formule fermée. La résolution de ce système passe par l'utilisation d'algorithmes d'optimisation tels que celui de Newton-Raphson, détaillé en annexes B.2.3 et B.2.4.

L'estimation de la probabilité individuelle de rachat découle directement des estimations des coefficients de régression par

$$\hat{p} = \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_k X_k), \tag{2.11}$$

où les $\hat{\beta}_i$ sont les coefficients de régression estimés par ML. Chaque assuré a donc sa propre probabilité de rachat, dépendante de ses caractéristiques personnelles et contractuelles. Il est capital d'avoir une idée de la précision de cette estimation. Pour ce faire, le calcul d'un intervalle de confiance est indispensable tant sur le plan individuel que sur le plan collectif (au niveau du portefeuille) lorsque nous voulons aggréger les résultats pour reconstruire le taux de rachat du portefeuille.

Plaçons nous dans le cadre collectif (à l'échelle du portefeuille). L'approximation normale de la loi binomiale pourrait constituer le point de départ de la construction de cet intervalle de confiance. Cependant, cette approximation requiert deux hypothèses sous-jacentes : i) $n \to \infty$ (grand nombre d'individus), ii) la probabilité p_i de rachat est comparable pour l'ensemble des individus (hypothèse d'homogénéité).
Le premier point n'est généralement pas un problème dans l'industrie de l'assurance (les portefeuilles sont souvent gros par nature). Le deuxième point est une condition nécessaire pour l'application du théorème de la limite centrale (TCL) : la somme de variables aléatoires i.i.d. suit une loi gaussienne. Le portefeuille est en réalité très hétérogène, mais on pourrait former i groupes d'assurés homogènes (en termes de caractéristiques), chacun de taille n_i. De plus, les assurés à l'intérieur de chaque groupe i sont considérés independants. Le nombre de rachats N_i^s du groupe i composé de n_i assurés est donc binomialement distribué (par propriété) ou normalement distribué (TCL, somme de Bernouilli i.i.d.), bien que l'hypothèse d'indépendance puisse paraître relativement discutable car l'environnement extérieur peut affecter un ensemble d'agents simultanément. Ainsi pour chaque groupe i,

$$\mathbb{E}[N_i^s] = \sum_{i=1}^{n_i} p_i = n_i p_i \ , \tag{2.12}$$

$$\mathbb{V}ar[N_i^s] = \sum_{i=1}^{n_i} p_i(1 - p_i) = \sum_{i=1}^{n_i} p_i q_i = n_i p_i q_i. \tag{2.13}$$

A partir de (2.12) et (2.13), nous obtenons l'intervalle de confiance de la probabilité de rachat $\hat{p}_i = N_i^s/n_i$ du i^{eme} groupe homogène en utilisant celui de la distribution

gaussienne. Le nombre total de rachats N^s du portefeuille est la somme des rachats de chaque sous-groupe homogène : $N^s = \sum_i N_i^s$. Or la loi Normale est stable par somme, donc N^s est toujours normalement distribué. Nous avons donc finalement une bonne approximation de la probabilité de rachat du portefeuille $\hat{p} = N^s/n$ par

$$\hat{p} = \sum_i N_i^s/n \sim N\Big(\frac{1}{n}\sum_i n_i p_i, \frac{1}{n^2}\sum_i n_i p_i(1-p_i)\Big),$$

qui conduit logiquement à l'intervalle de confiance (au niveau 5%)

$$[A - 1.96 \times B, A + 1.96 \times B] \tag{2.14}$$

où $A = \dfrac{\sum_i n_i p_i}{n}$, $B = \sqrt{\dfrac{\sum_i n_i p_i(1-p_i)}{n^2}}$, i est l'indice des sous-groupes homogènes, et p_i est la probabilité de rachat correspondante estimée.
Pour des soucis de concision, nous ne présentons pas ici les tests statistiques conduisant à la validation du modèle. Ces fameux tests du ratio de vraisemblance (pour la validation du modèle) et de Wald (pour la validation de chaque covariable) sont détaillés en annexe B.2.5.

2.2.3 Interprétations des résultats

Les valeurs estimées des coefficients de régression nous renseignent sur l'impact de chaque facteur de risque. L'ordonnée à l'oriine β_0 correspond à la valeur de z pour le profil de risque de référence : c'est la valeur moyenne de la réponse lorsque les covariables du prédicteur valent les modalités de référence pour les variables catégorielles, et sont nulles pour les variables continues (à condition d'avoir centré ces covariables continues en amont, sinon lorsqu'elles valent leur moyenne). Les coefficients β_i (i = 1,2,...,k) décrivent la contribution de chaque risque : un β_i positif signifie que si le facteur de risque augmente alors la probabilité de rachat augmente (corrélation positive), alors que s'il est négatif l'évolution se fait en sens inverse. Si la valeur absolue de $\beta_i/\sigma(\beta_i)$ (où $\sigma(\beta_i)$ est l'écart-type de l'estimation du coefficient) est grande, alors le facteur de risque i a une forte influence sur la probabilité de rachat, et inversement. Ces coefficients sont à comparer au profil de risque de référence, pour lequel $\beta = 0$ (sauf pour β_0).

Les professionnels aiment bien utiliser le rapport de côte (OR pour "odd-ratio"), qui expriment le rapport de probabilité $\frac{p}{1-p}$. Prenons un exemple d'illustration : la probabilité de rachat p=$P(Y$=$1|X)$ vaut 0,7. L'OR vaut donc p/q=0.7/0.3=2.33, ce qui veut dire qu'avec les mêmes caractéristiques X, le rachat a 2,33 fois plus de chance de se produire que le non-rachat. Cette idée se généralise lorsque les praticiens veulent évaluer la différence en termes de probabilité de rachat avec un changement des caractéristiques entre deux individus, prenons comme exemple l'âge. A partir de (2.11) nous avons pour un assuré donné $p/q = e^{\beta_0+\beta_1 X_{age}}$. Lors de la comparaison de deux individus ne différant que par leur âge (40 et 30 ans), tous les termes disparaissent excepté l'âge, l'OR vaut donc

$$\frac{P(Y=1|X_{age}=40)}{P(Y=0|X_{age}=40)} \Big/ \frac{P(Y=1|X_{age}=30)}{P(Y=0|X_{age}=30)} = \frac{e^{40\beta_1}}{e^{30\beta_1}} = e^{10\beta_1}$$

Nous constatons que la variation des valeurs de variables explicatives entraîne la multiplication de l'OR par des constantes liées aux coefficients de régression. Ces OR sont un outil opérationnel très utile (car simple) pour la définition de classe de risque.

2.2.4 Limites de la modélisation LR, améliorations

Les hypothèses formulées pour la mise en place de la modélisation constituent les principales limites du modèle. En particulier, les assurés sont considérés conditionnellement indépendants (formellement $Y_i|X_i$ sont indépendants) sachant leurs caractéristiques. Nous avons déjà évoqué le problème posé par cette hypothèse, qui n'est bien sûr pas totalement vérifiée en réalité. D'autre part l'estimation des paramètres du modèle est réalisable à condition que le coefficient de corrélation de Pearson ne vaille pas 100% dans l'étude des corrélations entre covariables, auquel cas la matrice des covariables (ou de "design") à inverser est singulière. La LR nécessite également un important volume de données afin d'en assurer la robustesse, mais cela ne semble pas être un écueil majeur en Assurance. L'utilisateur doit cependant s'assurer de la pertinence de ses données d'origine, point qui peut s'illustrer par l'exemple suivant : considérons un très ancien (disons 50 ans) portefeuille en *run-off* (pas de nouvelles affaires). Presque tous les assurés auront racheté leur contrat, la régression n'aurait plus beaucoup de sens en termes de segmentation (ce n'est pas le cas dans nos applications) et les résultats seraient probablement moins utiles. Enfin un seuil est à définir dans une problématique de prévision des classifications : une amélioration possible de la modélisation réside dans le choix optimal de ce seuil d'attribution de la réponse binaire. Le seuil naturel de 0,5 que nous considérons signifie qu'une probabilité prédite plus grande que 0,5 se verra attribuée la réponse 1 (rachat), sinon 0. Ruiz-Gazen & Villa (2007), Liu et al. (2006) et Lemmens & Croux (2006) montrent dans leurs articles que ce choix n'est pas optimal en cas d'échantillon avec réponses largement deséquilibrées. Pour éviter d'obtenir des résultats non-représentatifs de la réalité, leurs méthodes se basent sur des techniques de reéchantillonage de type *importance sampling*.

2.3 Illustration : application sur des contrats mixtes

Les informations dont nous disposons dépendent de l'entité pays d'AXA qui nous les fournit. La majorité des bases de données comprennent la date de naissance des assurés, leur sexe, leur lieu de résidence, la date d'émission du contrat, la date de fin du contrat, le type de contrat, la fréquence de la prime, la somme assurée. Nous avons aussi dans certains cas des renseignements concernant le statut marital, le statut fumeur ou encore le réseau de distribution (cette liste ne se veut pas exhaustive). Dans cette partie, nous nous focalisons sur le portefeuille espagnol d'AXA et plus précisément les contrats mixtes. **Précisons d'ailleurs que l'ensemble des études de cas des chapitres 2, 3 et 4 sont réalisées avec ces mêmes contrats mixtes pour favoriser une certaine continuité**.

Nous essayons de segmenter notre population d'assurés en fonction de leur type de contrat, de leur sexe, de leur âge, de leur fréquence de prime, de leur somme assurée, et du montant de leur prime. La somme assurée (notée "face amount" dans la base de données) est un indicateur de la richesse de l'assuré, la prime contient elle la prime de risque (liée à la garantie du risque Vie considéré, ici le décès) et la prime d'épargne. La prime de risque est le produit actualisé de la somme sous-risque par la probabilité de déclencher la garantie. La prime d'épargne est l'investissement de l'assuré. Nous utilisons en partie la librairie `rpart` de `R` pour implémenter la méthode CART et obtenir nos résultats. Les fonctions servant à résoudre le problème d'optimisation dans l'estimation du modèle LR sont déjà intégrées au coeur des programmes `R`.

Analyse statique Nous entendons par analyse statique une photographie de l'état du portefeuille en décembre 2007. Les contrats de cette base de données sont des contrats de pure épargne et des contrats mixtes. Les résultats numériques ci-après concernent l'étude des 28506 contrats mixtes, isolés des contrats pure épargne car les comportements de rachat doivent être distingués par grande ligne de produit. Ces contrats annuellement renouvelables (TAR) ne peuvent pas être rachetés avant un an d'ancienneté (sauf cas exceptionnel), et il est à noter qu'il n'y a pas de dispositif fiscal particulier en Espagne sur les contrats d'Assurance Vie avec composante épargne. Cela signifie ici que les assurés peuvent racheter leur contrat à chaque date anniversaire sans frais, mais sont pénalisés en cas de rachat à un autre moment. Nous voyons en figure 2.3 que ceci est une incitation importante qui dicte le profil des rachats en fonction de l'ancienneté du contrat.

L'étude couvre la période 2000-2007, les caractéristiques des assurés et de leur contrat sont celles observées soit à la date de rachat, soit en Décembre 2007 (si pas de rachat). Rappelons que notre but est de trouver les principaux déclencheurs de rachat en se servant des variables explicatives observées, ce qui nous permettra de détecter des agents "risqués" en termes de décision de rachat **à une date donnée**. Certaines précautions sont à prendre dans ce type d'analyse faute de quoi les résultats peuvent être particulièrement biaisés. Nous pensons à la composition du portefeuille, à son degré de maturité, à la part des nouvelles affaires. Par exemple, si les rachats ne sont observés qu'après une certaine ancienneté, il est important de s'assurer que le portefeuille soit à maturité (sinon nous observerions un taux de rachat quasi-nul, ce qui nous amènerait à des conclusions erronées). Nous verrons que c'est l'un des gros défauts de l'analyse statique qui peut introduire un biais important, par opposition à l'analyse dynamique du chapitre 3 qui tente de corriger ceci.
En Décembre 2007, 15571 des 28506 contrats mixtes ont été rachetés, soit environ 55%. Les deux modèles de segmentation présentés nous apportent deux informations complémentaires :

- CART nous donne les variables les plus discriminantes par rapport aux comportements de rachat en ordre décroissant (en lisant l'arbre depuis la racine jusqu'aux feuilles). Au final, nous classons un assuré comme "risqué" à l'aide d'une prévision binaire (bien qu'on puisse accéder aux probabilités précises de chaque classe et donc à la probabilité de rachat en l'occurence) ;
- LR offre un résultat plus numérique : la probabilité de racheter son contrat étant donné ses caractéristiques et la sensibilité de sa décision aux évolutions des covariables (avec les OR).

2.3.1 Résultats par les CART

Nous réalisons l'analyse sous `R` grâce à la librairie `rpart` *, et plus précisément par la procédure `rpart` qui construit l'arbre de classification. Par défaut `rpart` utilise l'indice de Gini pour calculer l'impureté d'un noeud, mais cette option n'est pas très importante puisque les résultats ne sont quasiment pas impactés. Il n'y a pas de coût de mauvaise classification introduit (voir annexe B.1.3) dans notre application.
Nous procédons comme en théorie et construisons d'abord l'arbre T_{max} sans coût de complexité (en posant l'option *cp* égale à 0), puis l'arbre est élagué ("pruned tree") pour diminuer son nombre de feuilles et simplifier les résultats. Le nombre minimal

∗. r-partitionning : http ://cran.r-project.org/web/packages/rpart/index.html

d'observations dans une feuille a été fixé à 1, le nombre de divisions concurrentes calculées est de 2. Nous créons aléatoirement les échantillons d'apprentissage et de validation, dont les tailles respectives sont de 16868 et 11638 assurés.

L'estimation par échantillon témoin de l'erreur de prévision de l'arbre maximal T_{max} calculée sur l'échantillon de validation est de 14.88%, correspondant aux termes non-diagonaux de la matrice de confusion du tableau 2.1. Cet arbre a trop de feuilles et admet une représentation trop complexe, ce à quoi nous remédions en l'élaguant. Le choix du paramètre de complexité α lors de l'élagage (annexe B.1.3) est un arbitrage entre la taille finale de l'arbre et le taux de mauvaise classification désiré par l'utilisateur. Le tableau B.1 et la figure B.3 en annexe B.1.2 trace l'erreur d'apprentissage en fonction du coût de complexité. Dans ce graphe, à chaque paramètre de complexité est associé un arbre optimal dont la taille est donnée, ce qui permet de choisir le α optimal par minimisation de l'erreur d'apprentissage. Nous obtenons $\alpha \in]1.04e^{-04}, 1.30e^{-04}]$, mais le nombre de feuilles correspondant (82) est encore trop élevé à notre goût. Nous avons donc choisi $\alpha = 6e^{-04}$, ce qui correspond à 11 feuilles pour une très faible perte de précision dans la classification. Cet arbre est visualisable en figure 2.1. La variable la plus discriminante semble toujours être le type de contrat (caractérisé en fait par le type de prime, unique ou périodique ; et l'option de participation au bénéfice), puis l'ancienneté du contrat et ainsi de suite.

Les variables sélectionnées dans la construction de l'arbre sont le type de contrat, l'ancienneté, la richesse de l'assuré ("face amount"), la fréquence de prime, la prime d'épargne et l'âge de souscription. Nous remarquons que le sexe et la prime de risque n'apparaissent pas dans cet arbre final, parce que leurs effets ne semblent pas être significatifs. La première règle de division est "L'assuré possède t'il l'option de participation au bénéfice ?". Si la réponse est "non", alors descendre dans la branche de gauche,

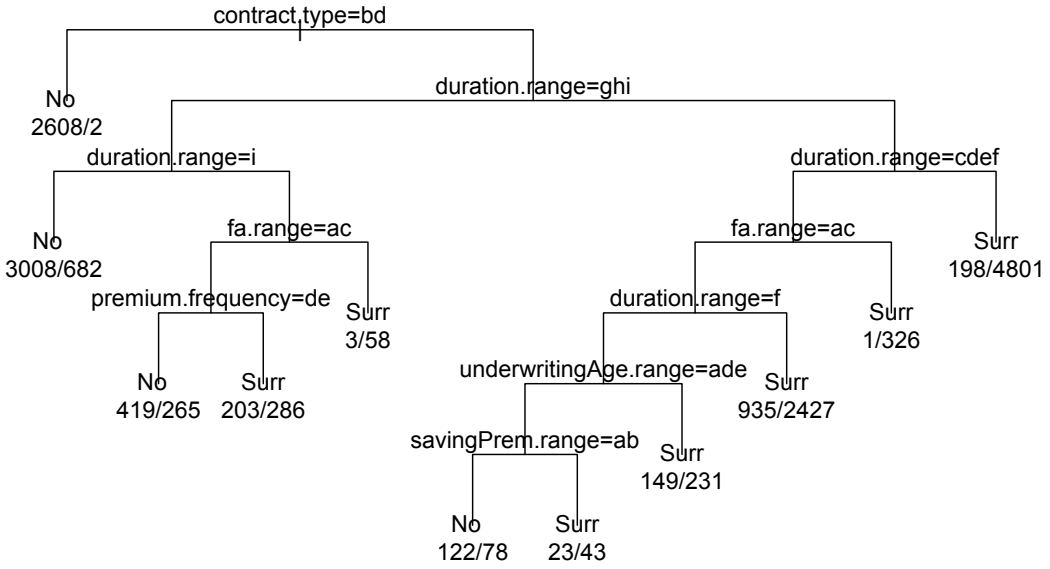


FIGURE 2.1 – L'arbre final de classification. Variable réponse binaire : rachat. La première règle de division *contract.type = bd* signifie que le type de contrat est la variable la plus discriminante (*bd* correspond aux 2^{eme} et 4^{eme} modalités, comme dans l'ordre alphabétique). Les variables continues ont été catégorisées pour la modélisation.

TABLE 2.1 – Matrice de confusion (T_{max}), échantillon de validation.

	observed Y = 0	observed Y = 1
predicted Y = 0	4262	1004
predicted Y = 1	728	5644

TABLE 2.2 – Matrice de confusion (arbre élagué), échantillon de validation.

	observed Y = 0	observed Y = 1
predicted Y = 0	4188	1078
predicted Y = 1	664	5708

sinon descendre dans la branche de droite. Les classes prédites (rachat ou non-rachat) sont écrites sur les feuilles, les proportions qui y apparaissent sont le nombre d'assurés n'ayant pas racheté versus ceux qui ont racheté leur contrat. Plus la différence entre ces deux nombres est grande, meilleure est la segmentation. Ici, un assuré dont le contrat ne contient pas l'option de participation au bénéfice a 99,92% (2608/2610) de ne pas racheter, quelque soit le format de sa prime (PP sin PB et PU sin PB, voir légende du tableau 2.4). La classe attribuée est donc "No", équivalente à "pas de rachat". Considérons un assuré dont les caractéristiques sont une prime périodique, un contrat avec clause de participation au bénéfice. Son ancienneté appartient aujourd'hui à la septième modalité de la variable catégorisée, et sa richesse se situe dans la deuxième classe. La prévision de l'arbre est que cet assuré aura 95% (58/61) de chance de racheter, cet assuré est donc considéré comme risqué.

Il est évident que le facteur de risque le plus discriminant lorsque nous regardons la figure 2.1 est l'option de participation au bénéfice. Le taux de mauvaise classification (erreur d'apprentissage) de cet arbre est de 15% (33.1% × 45.4%, où 45.4% est l'erreur de la racine quand aucune division n'est réalisée) d'après le tableau B.1 des erreurs relatives de l'annexe B.1.2. L'erreur de prévision de 14.97% peut être estimée via la matrice de

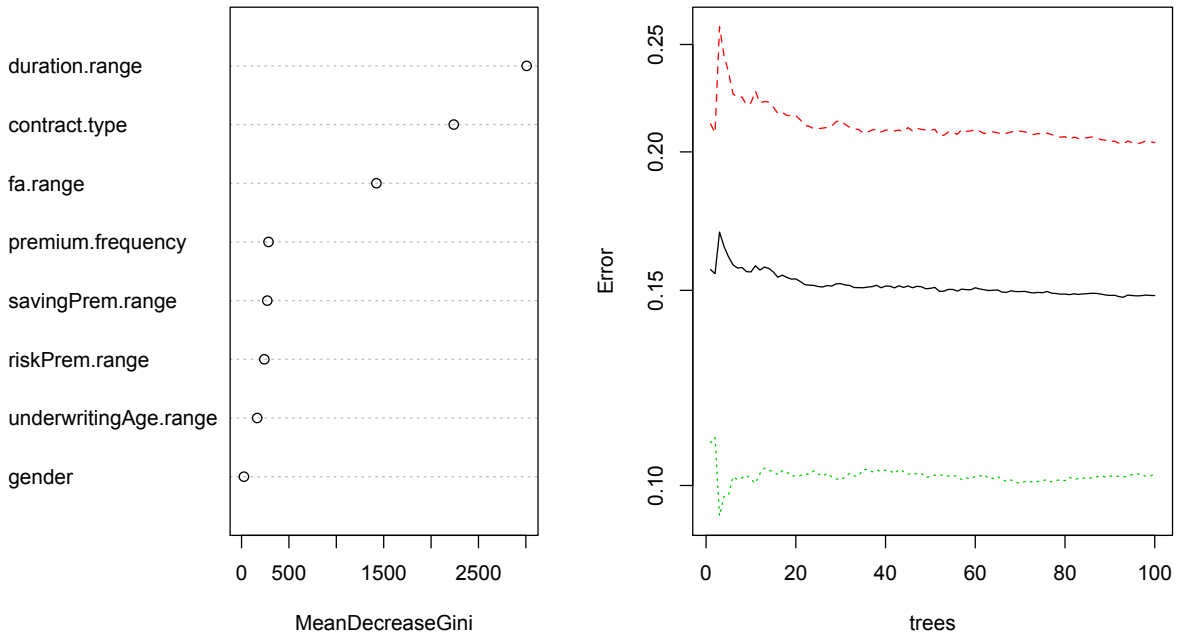


FIGURE 2.2 – Sur la gauche : l'importance des variables explicatives. Sur la droite : le nombre d'arbres requis pour stabiliser l'erreur *out-of-bag* : la courbe noire est l'erreur globale, la verte est l'erreur pour la réponse "rachat" et la rouge l'erreur pour la réponse "pas de rachat".

TABLE 2.3 – Matrice de confusion du classifieur obtenu par les forêts aléatoires.

	observés Y = 0	observés Y = 1
prédits Y = 0	10327	2608
prédits Y = 1	1592	13979

confusion du tableau 2.2 et est relativement satisfaisante puisqu'elle reste proche de l'erreur de prévision de l'arbre maximal T_{max}. Nous en déduisons que le compromis est très intéressant : l'élagage d'un arbre de 175 feuilles à un arbre de 11 feuilles nous fait perdre moins de 1% d'erreur de prévision !

Nous utilisons les *bagging predictors* pour consolider ces résultats avec la librairie `randomForest`. Lors de l'utilisation de l'algorithme des forêts aléatoires, les étapes successives permettent de construire un ensemble de classifieurs bootstrap. L'aggrégation de ces classifieurs amène à un classifieur final, qui ne peut cependant pas être représenté sous forme d'arbre mais qui fournit des résultats plus robustes (tous les concepts utilisés dans cet algorithme sont consultables sur la page web de Breiman ∗). Le tableau 2.3 résume les résultats de classification sur l'échantillon d'origine (pas d'échantillon d'apprentissage ni de validation car il s'agit déjà de méthodes bootstrap) : l'estimation de l'erreur sans biais appelée erreur *out-of-bag* est de 14.73%. L'importance des variables explicatives dans le processus de classification est visualisable en figure 2.2, de même que le nombre d'arbres nécessaires dans la forêt pour la stabilisation de l'erreur *out-of-bag* (environ 50 arbres ici). Ces résultats viennent confirmer nos attentes : l'ancienneté du contrat et son type sont encore une fois les variables les plus discriminantes pour expliquer les décisions de rachat des assurés. Afin de s'assurer que l'effet de l'ancienneté du contrat (effet temporel) ne biaise pas ces résultats, nous avons décidé de relancer l'analyse en isolant cet effet par une séparation dans notre jeu de données : d'un côté les personnes dont le rachat s'est effectué pour des anciennetés de contrat correspondantes aux pics observés en

∗. See http ://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm

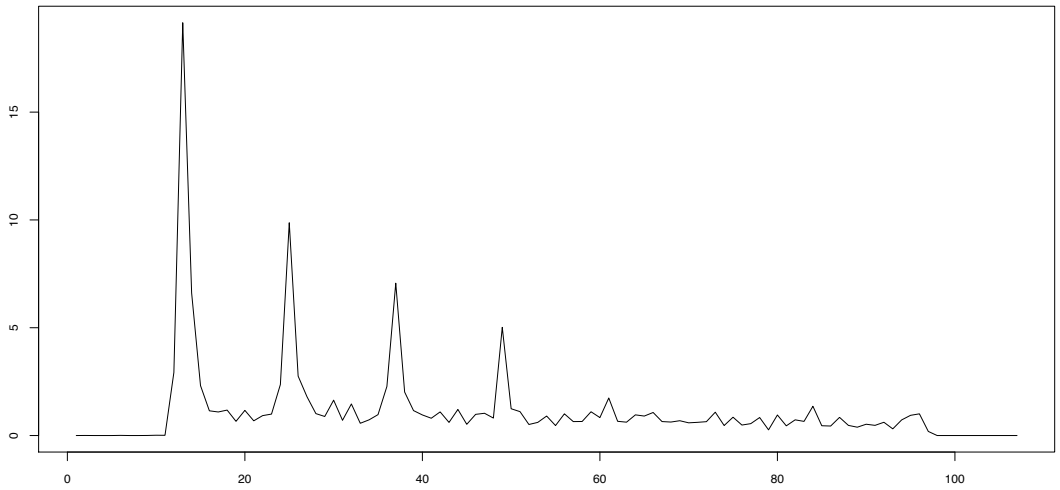


FIGURE 2.3 – Taux de rachat (%) VS ancienneté (en mois) pour les contrats Mixtes. Effet des pénalités de rachat (les contrats peuvent être rachetés sans frais à chaque date anniversaire, ce qui explique les pics de rachat observés).

FIGURE 2.4 – Importance des variables explicatives en excluant l'effet de l'ancienneté. Sur la gauche les assurés dont l'ancienneté du contrat correspond aux pics observés en Figure 2.3, et autres assurés sur la droite.

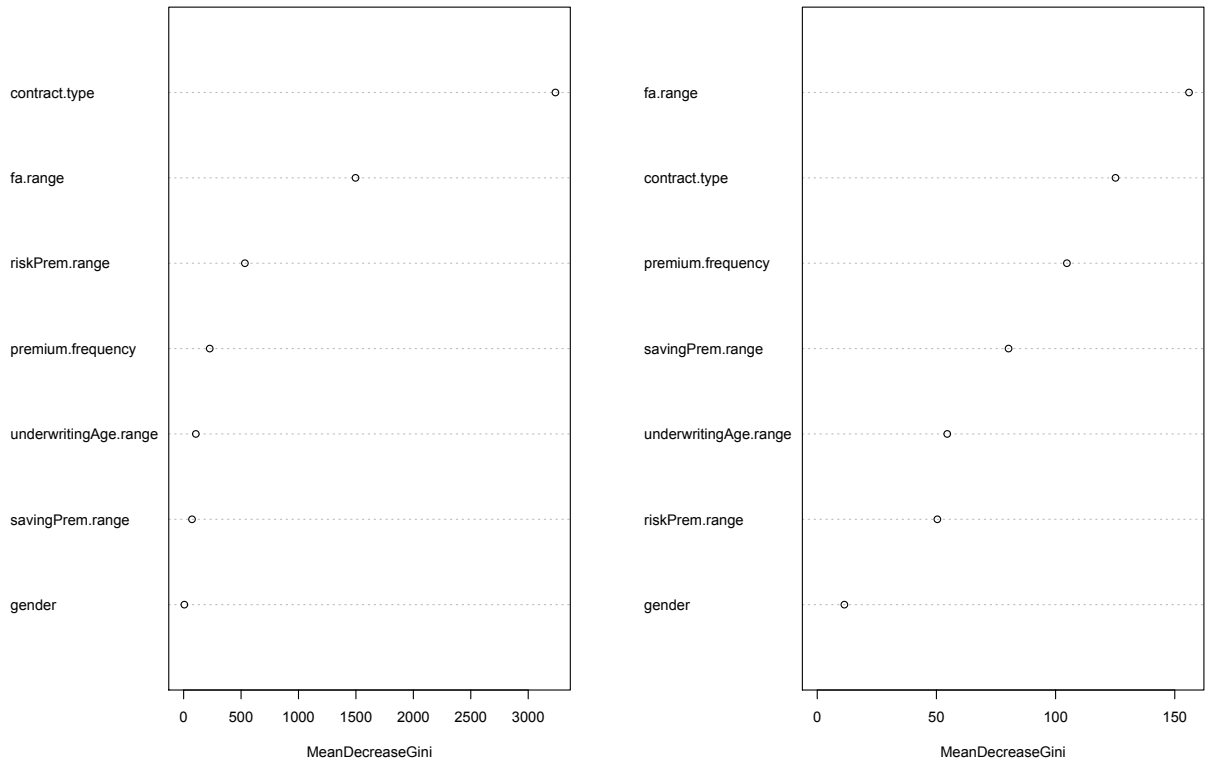


figure 2.3, de l'autre les assurés restants. Nous regardons ainsi les rachats provoqués par les contraintes de frais à payer, mais aussi ceux qui ne le sont pas. La figure 2.4 montre que les facteurs discriminants principaux restent les mêmes quelque soit la population étudiée (l'ordre diffère légèrement), ce qui signifie que l'effet de l'ancienneté n'est pas corrélé à un autre facteur de risque et n'introduit pas de biais dans les résultats que nous obtenons.

2.3.2 Classification par le modèle logistique (LR)

Le logiciel `R` et sa fonction `glm` nous permettent d'appliquer le modèle logistique à nos données. Comme détaillé dans la partie théorique, les sorties du modèle sont l'effet de chaque covariable (facteur) par les coefficients de régression, l'écart-type de l'estimation de ces coefficients, et la déviance du modèle (cf annexes B.2.3, B.2.4 et B.2.5).

Lors de la résolution du système d'équation amenant à l'estimation des coefficients de régression, les variables catégorielles sont introduites par une suite de variables indicatrices (une par modalité) qui permet de définir la matrice de "design" qui sera inversée par la procédure `glm`. Cette fonction utilise un algorithme itératif pas-à-pas dans le but de comparer un modèle basé sur p' des p variables d'origine à n'importe quel sous-modèle (contenant une variable de moins), ou même à n'importe quel sur-modèle (avec une variable supplémentaire). La fonction `stepAIC` de la librairie `MASS` nous permet de sélectionner de manière adéquate les variables pertinentes, pour finalement obtenir un modèle optimal qui contient un minimum de variables explicatives pertinentes. A

TABLE 2.4 – Rapports de côte (OR), contrats mixtes (ancienneté en mois, échantillon d'apprentissage). Types de contrat : PP con PB → prime périodique (PP) avec participation au bénéfice (PB), PP sin PB → PP sans PB, PU con PB → prime unique (PU) avec PB, PU sin PB → PU sans PB. Les variables continues (ex : ancienneté) ont été catégorisées.

OR	Référence	Autres modalités							
Duration	[0,12]	]12,18]	]18,24]	]24,30]	]30,36]	]36,42]	]42,48]	]48,54]	> 54
Rachats	3062	1740	1187	791	728	400	365	244	682
OR empirique		10.56	2.89	2.69	1.82	1.16	0.96	0.68	0.19
OR modélisé		0.27	0.07	0.06	0.05	0.03	0.02	0.02	0.004
Fréquence de prime	Mensuelle	Bi-mensuelle	Trimestrielle	Semestrielle	Annuelle	Unique			
Rachats	2790	12	323	92	595	5387			
OR empirique		2.22	0.93	0.66	2.39	1.60			
OR modélisé		2.52	0.97	0.80	1.55	0.75			
Age souscription	[0,20[	[20,30[	[30,40[	[40,50[	[50,60[	[60,70[	> 70		
Rachats	258	1719	2165	2002	1490	1088	477		
OR empirique		1.16	1.06	1.25	1.63	2.67	3.28		
OR modélisé		1.32	0.99	0.77	0.67	0.51	0.47		
Face amount	#1	#2	#3						
Rachats	5361	684	3154						
OR empirique		0.14	0.12						
OR modélisé		0.003	0.0008						
Prime de risque	#1	#2	#3						
Rachats	3941	2987	2271						
OR empirique		1.50	0.92						
OR modélisé		1.43	1.30						
Prime d'épargne	#1	#2	#3						
Rachats	3331	1762	4106						
OR empirique		1.90	2.09						
OR modélisé		2.55	3.78						
Type de contrat	PP con PB	PP sin PB	PU con PB	PU sin PB					
Rachats	3840	0	5357	2					
OR empirique		0	4.75	0.0008					
OR modélisé		5.6e-08	0.0006	3.9e-06					

des fins de comparaison, les échantillons d'apprentissage et de validation sont les mêmes que dans l'application CART. Comme d'habitude les coefficients de régression ont été estimés sur l'échantillon d'apprentissage alors que les prévisions se sont effectuées sur l'échantillon de validation. Le tableau B.2 en annexe B.2.1 récapitule l'ensemble des résultats, à savoir les coefficients de régression et leur écart-type, les p-valeurs des tests de Wald (confiance dans l'estimation et significativité des coefficients, cf annexe B.2.5). Nous déduisons de ce tableau que les covariables qui semblent avoir le plus d'impact (grande valeur absolue) sont encore une fois l'ancienneté du contrat, le type de contrat, mais aussi la richesse de l'assuré ; ce qui est en ligne avec les résultats obtenus par la méthode CART. Le fait que l'ancienneté de contrat soit un facteur explicatif clef rend très important le profil des rachats en fonction de celle-ci, et notamment la prise en compte de changement de législation par exemple (taxation, fiscalité...).

Les rapports de côte (OR) présentés en section 2.2.3 sont à comparer à la valeur 1 (modalité de référence). Nous constatons au regard du tableau 2.4 que les OR modélisés représentent assez mal la réalité car de grosses différences existent avec les OR empiriques (obtenus par statistiques descriptives). Par exemple, le modèle prévoit qu'une personne âgée de 70 ans ait moins de chance de racheter qu'un jeune assuré de moins de 20 ans, toutes caractéristiques égales par ailleurs. L'expérience montre qu'ils sont en fait 3,28 fois plus susceptibles de racheter ! Nous retenons donc de ce tableau que les

TABLE 2.5 – Matrice de confusion (modèle LR).

	observé Y = 0	observé Y = 1
prédit Y = 0	#correct rejections 4153	#misses 637
prédit Y = 1	#false risky policyholder 1113	#success 5735

TABLE 2.6 – Critères de performance.

	T_{max}	T_{pruned}	$T_{RandomForest}$	LR
Se	84.9%	84.1%	84.3%	90%
Sp	85.4%	86.3%	86.7%	78.9%
(1-Se)	15.1%	15.9%	15.7%	10%

OR estimés varient très souvent dans le même sens que les OR observés. C'est le cas si l'on considère le facteur d'ancienneté : la figure 2.3 expose le profil des rachats en fonction de l'ancienneté (pourcentage des rachats pour chaque tranche d'ancienneté) et confirme les estimations des OR (tableau 2.4) liés à ce facteur : effectivement le risque est important à partir de la première date anniversaire et décroît dans le temps.

Le modèle a globalement une mauvaise qualité d'ajustement au regard de la significativité des estimations des coefficients de régression, c'est d'ailleurs ce qui explique que les OR empiriques et modélisés soient si éloignés. Toutefois, il faut garder en tête que notre problématique initiale est de segmenter notre population d'assuré et de faire des prévisions de profil de risque. Nous préférons donc privilégier le pouvoir prédictif à la qualité d'ajustement dans l'arbitrage naturel entre ces deux notions. La matrice de confusion du tableau 2.5 fournit le nombre d'assurés mal classés et représente le pouvoir prédictif de cette méthode, la lecture de cette table se faisant de la même manière qu'avec les CART. Une hypothèse supplémentaire est utilisée ici, à savoir que nous attribuons une réponse de rachat lorsque la probabilité de rachat modélisée excède 0,5 et inversement. Les bonnes prévisions représentent 84.96% de l'échantillon de validation ; ce qui donne une erreur de prévision de 15.04%, un résultat quasi-similaire à celui obtenu par l'algorithme CART.

D'autres critères, couramment appelés critères de performance, servent à comparer les classifieurs : il s'agit de la sensibilité (Se) et de la spécificité (Sp). Appelons *success* la case correspondante à une réponse observée et prédite de rachat dans la matrice de confusion. Les *misses* correspondent à une réponse prédite de non-rachat tandis que l'observation était un rachat. Les *correct rejections* correspondent à une réponse observée et prédite de non-rachat, enfin les *false risky policyholder* désignent une réponse prédite de rachat alors qu'il n'y en pas eu dans la réalité. La sensibilité est définie comme le nombre de *success* sur le nombre de contrats rachetés observés, et la spécificité est le nombre de *correct rejections* sur le nombre de contrats non-rachetés observés. Le tableau 2.6 résume les critères de performance des différentes méthodes de classification ; sachant que ce qui nous intéresse avant tout dans ce contexte est de minimiser les *misses*. Les prévisions par la LR présentent moins de *misses* et plus de *false risky policyholders*, les résultats étant comparables et les erreurs équilibrées entre les trois applications de CART. Le compromis entre sensibilité et spécificité est meilleur avec CART mais le nombre de *misses* est plus élevé, ce qui nous conduirait ici à choisir le modèle LR ici (10%) pour plus de prudence.

2.4 Conclusion

L'objectif de ce chapitre était de présenter deux modèles de segmentation qui apportent des réponses sur le profil de risque des assurés, par la prise en compte de leurs caractéristiques individuelles et des options de leurs contrats. Qu'avons nous appris ?

Cette étude a permis de mettre en exergue quelques types de profils risqués : **les personnes jeune ont tendance à racheter davantage que les autres**, comme ceux qui ont une **prime périodique** ("annuelle" et "bi-mensuelle" sont les pires cas). Les assurés les plus pauvres (au vu de l'indicateur "face amount" bien sûr, ce qui ne veut pas dire forcément qu'ils le sont) rachèteront leur contrat probablement plus souvent : en effet ils doivent payer des frais et des primes régulières mais n'ont pas l'argent pour, alors que les personnes plus riches n'y prêtent pas vraiment attention. La majeure concentration du risque se situe grosso modo sur les premiers jours (premières semaines) qui suivent la levée d'une contrainte fiscale ou d'une contrainte prévue par le contrat : **lorsque l'ancienneté atteint ce seuil, le risque est très élevé**. Dans une optique de segmentation de risque à la souscription, notons que ce facteur de risque est une information inexistante qui ne peut donc pas être prise, ce qui justifie le fait que nous ayons regardé le classement par importance des facteurs de risque en isolant cet effet. Enfin, la clause de participation aux bénéfices (PB) de l'entreprise pour l'assuré semble jouer un rôle clef dans le processus de décision du rachat, l'étude ayant montré que les personnes **sans cette option ne rachètent que très peu** leur contrat alors que les autres le font tôt ou tard. Trois principales raisons pourraient expliquer ce phénomène : premièrement les agents rachètent pour basculer leur épargne sur un nouveau produit offrant un taux de PB supérieur au leur, deuxièmement un taux de PB attractif pendant les premières années du contrat permet à l'assuré de surperformer le rendement initial et l'inciter à racheter par la suite dans le but de récupérer une bonne valeur de rachat, troisièmement le simple fait de recevoir annuellement l'information sur le taux de PB qui sera versé par le contrat et la valeur de rachat associée peut jouer sur une décision de rachat. Le **sexe de l'assuré n'apparaît pas comme un facteur de risque à prendre en compte**.

D'un point de vue plus technique, nous avons vu que le processus de classification peut se réaliser soit par l'emploi du modèle logistique soit par l'emploi des méthodes CART, les résultats étant en adéquation. Des profils type de risque se dégagent plus facilement à partir de statistiques descriptives ou des CART, tandis que le modèle LR donnent accès à des indicateurs intéressants tels que les rapports de côte. Les deux modèles apportent des résultats complémentaires et font intervenir des hypothèses bien différentes, mais servent globalement une même cause : une réduction de dimension de l'espace des données, autrement dit une sélection des variables les plus discriminantes en termes de rachat (dans le but de simplifier la future modélisation). Un outil informatique (basé sur RExcel) permettant d'obtenir un large panel de statistiques descriptives ainsi que l'usage de ces deux modèles de segmentation a été implémenté à cette occasion pour étudier les comportements de rachat à plusieurs niveaux d'échelle (numéro de produit, ligne de produit, famille de produit, pays) dans quatre entités d'AXA (Espagne, Etats-Unis, Belgique, Suisse).

Enfin, cette analyse statique est utile pour comprendre quelles sont les caractéristiques des contrats et des assurés qui ont un rôle dans les décisions de rachat. Elle présente l'inconvénient majeur qu'elle ne tient pas compte de l'impact du contexte économique et financier sur les comportements de rachat puisque nous regardons l'état du porte-

feuille à une date fixée (fin 2007), supprimant ainsi les effets temporels. Nous pourrions argumenter que dans un contexte économique classique les comportements de rachat ne sont pas guidés par celui-ci, et donc que cette analyse suffit. Cependant, lorsque l'environnement (économie, image de la compagnie) change, il devient très difficile d'anticiper les comportements de rachat pour reconstruire le taux de rachat à l'échelle du portefeuille. La corrélation entre les décisions des agents est par exemple susceptible de fortement augmenter (Loisel & Milhaud (2011)). La modélisation devient de ce fait beaucoup plus compliquée et nous verrons la nécessité de bien capturer les effets endogènes (idiosynchratiques, structurels) **et** exogènes (conjoncturels) à travers les problèmes rencontrés dans le chapitre suivant lors de l'utilisation "dynamique" des GLM.

Chapitre 3

Crises de corrélation

Pouvoir recomposer précisément par date le taux de rachat à l'échelle d'un portefeuille d'assurance nécessite de conserver les caractéristiques individuelles des contrats et des assurés car nous avons vu que certaines variables étaient très discriminantes. Par conséquent l'évolution de la composition du portefeuille est importante pour une modélisation adéquate du taux de rachat : en effet si le profil du portefeuille d'assurance change beaucoup entre deux dates données, les décisions de rachat et donc le taux seront radicalement différents. Cette suggestion nous amène à retenir une fois de plus l'usage de modèles de régression qui permettent d'intégrer ces considérations par l'intermédiaire de covariables. Nous verrons également dans ce chapitre qu'une modélisation logistique classique est insuffisante, ce qui nous permettra d'introduire le phénomène de corrélation entre comportements et de présenter qualitativement son impact potentiel.

3.1 Problème de la régression logistique dynamique

Comme discuté auparavant, faire des prévisions de taux de rachat en se basant sur une analyse statique peut entraîner des erreurs importantes. Les modèles développés dans le chapitre précédent sont adaptés dans une optique de définition de classe (profil) de risque, mais il faut relativiser la pertinence de leur utilisation lors de l'étude d'un phénomène dont la modélisation dépend fortement d'un environnement mouvant. Dans ce cas, une analyse dynamique permet de prendre en compte les facteurs extérieurs et de mieux refléter l'influence de ceux-ci sur les décisions des assurés. Nous modélisons dans la suite le taux de rachat du portefeuille sur un pas de temps mensuel par aggrégation de décisions individuelles des assurés. Ces décisions sont retournées par le modèle logistique auquel nous ajoutons en variables explicatives des facteurs de risque économiques et financiers.

Hormis la prise en compte du contexte économique, cette analyse dynamique permet d'éviter certains problèmes évoqués au chapitre 2 comme la durée de la période couverte pour étudier le phénomène. Si elle est trop courte, on n'observerait aucun rachat (puisqu'il faut généralement un minimum d'ancienneté avant que le rachat puisse se faire), si elle est trop longue on observerait un taux de rachat proche de 100% ; les deux situations n'étant pas réalistes. De plus, le fait de modéliser mensuellement les décisions des assurés permet de rendre compte du fait que dans la réalité les assurés sont susceptibles de se poser fréquemment la question du rachat de leur contrat.
Grossièrement, cette modélisation dynamique pose deux problèmes majeurs : la stabilité et la robustesse. Ces écueils sont dûs à l'ajout d'une hypothèse très forte, l'indépendance temporelle entre les décisions. Nous considérons que la décision d'un assuré à la date

$t+1$ est indépendante de ce qui s'est passé avant, et notamment indépendante de sa décision à la date t. Pour construire la base de données nécessaire à cette analyse, nous dupliquons chaque assuré chaque mois où il est présent en portefeuille (ce qui nous donne un échantillon global de 991 010 lignes) et mettons à jour ses caractéristiques (indices économiques, ancienneté,...), à partir de la même base de données que celle utilisée au chapitre 2. Cette opération peut donner lieu à l'introduction d'un nouveau biais : les caractéristiques des assurés qui restent le plus longtemps en portefeuille sont sur-représentées. Ceci dit et après vérification, ce biais ne joue pas beaucoup sur nos résultats lorsque nous les comparons aux coefficients de régression d'un modèle de Cox où l'on ne duplique pas les assurés (car c'est un modèle de survie).

Après s'être assuré que nous lançons cette analyse sur une période représentative du portefeuille (à maturité), il est possible de voir la qualité de la modélisation en comparant le taux de rachat observé avec le taux de rachat prédit sur un pas mensuel. Les échantillons d'apprentissage et de validation sont construits différemment ici : l'apprentissage représente environ deux tiers de la période étudiée, soit de Janvier 2000 à Mars 2005 (629 357 observations) ; tandis que la validation s'effectue sur la période restante (Avril 2005 à Décembre 2007, 361 653 observations). Cette technique de validation "temporelle" permet de rendre compte de l'exposition des assurés à des contextes économiques différents, et ainsi de tester non pas uniquement la qualité d'adéquation du modèle mais aussi son réel pouvoir prédictif pour des simulations futures de taux de rachat. Les covariables introduites dans la régression logistique sont le mois d'observation (effet de saisonnalité) et le contexte économique (taux de chômage, taux crédité des contrats, Ibex 35, taux d'intérêt court terme 1Y et long terme 10Y *), en plus des covariables considérées dans l'analyse statique. Nous négligeons le décès des assurés lorsque nous réalisons les prévisions futures car c'est d'un évènement rare (taux d'environ $2e^{-4}$). La période d'observation a visiblement une grande influence

∗. Données économiques et financières récupérées sur les sites Yahoo Finance et OCDE Stats.

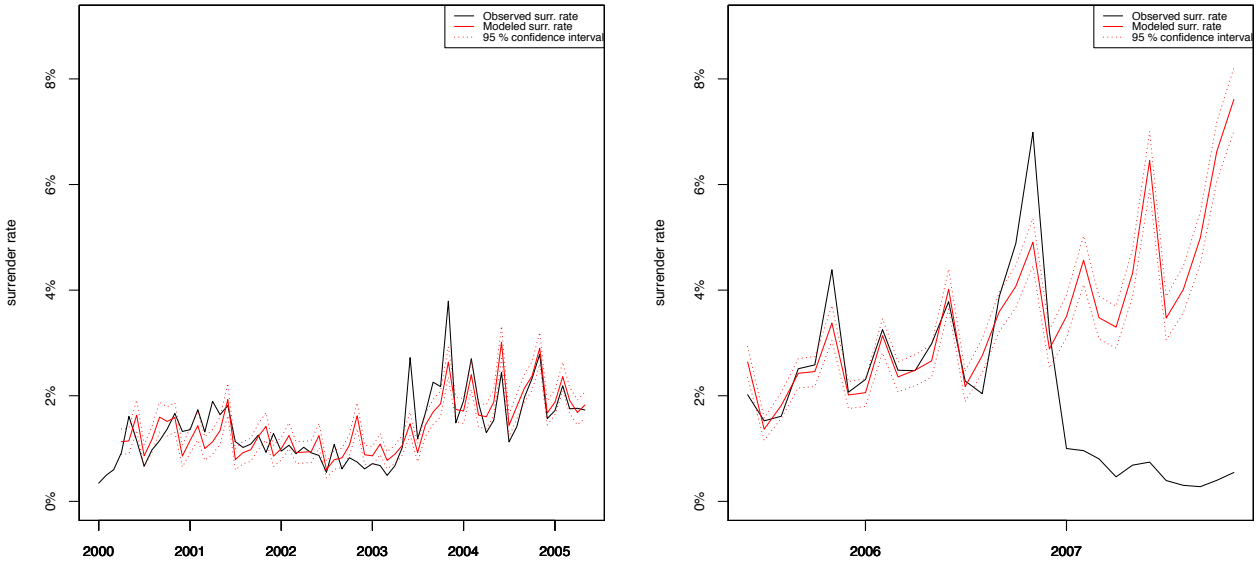
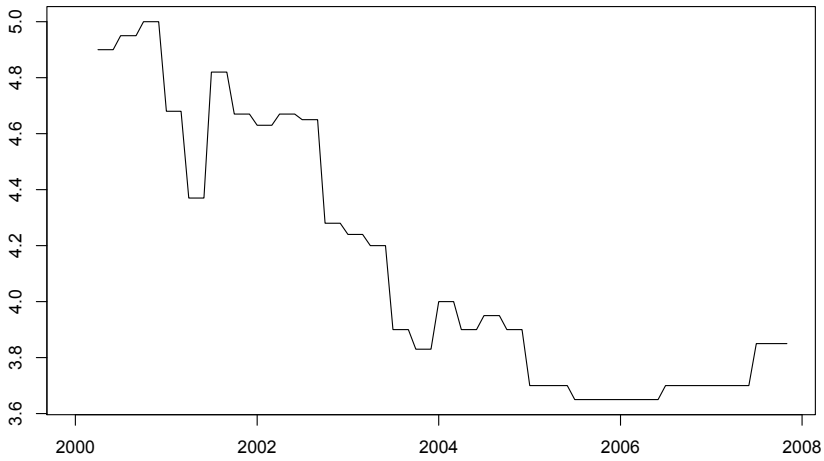


FIGURE 3.1 – Prévisions du taux de rachat collectif (du portefeuille) avec l'inclusion de covariables économiques. Sur la gauche, les prévisions sur l'échantillon d'apprentissage et sur la droite les prévisions sur l'échantillon de validation.

FIGURE 3.2 – Taux crédité mensuel des contrats Mixtes. Ce taux comprend le taux moyen garanti ainsi que taux de PB moyen servi.



sur le calibrage du modèle au vu des résultats de la figure 3.1 : nous constatons les bonnes qualités d'adéquation du modèle sur la période d'apprentissage, de même que ses mauvaises qualités prédictives observées en 2007. Les figures 3.1 et 3.2 montrent que le niveau moyen de rachat augmente lorsque le taux de participation aux bénéfices décroît fortement (2003-2004), dénotant une relation claire entre les taux crédités et les taux de rachat. Finalement les résultats ont l'air acceptable bien que le modèle marche très mal en situation extrême ; mais un modèle est-il censé marcher en régime extrême ? La crise financière qui s'est déclarée dans l'année 2007 a très certainement fait évoluer l'importance des facteurs explicatifs clef du rachat en accordant plus d'importance aux facteurs exogènes (indices boursiers, taux d'intérêts) qu'endogènes, ce qui ne semble pas forcément être le cas en régime de croisière. En période de crise, l'hypothèse d'indépendance (temporelle et entre agents) semble violée, ce qui fait chuter très nettement le taux de rachat du portefeuille sur les produits mixtes espagnols dans l'année 2007 : concrètement les taux garantis par les contrats sont très intéressants comparés aux taux d'intérêts qui baissent sans arrêt, incitant les assurés à prendre une décision commune (garder à tout prix leur contrat). Le modèle ne prévoit pas cette chute subite car il ne capte visiblement pas les effets dans leurs bonnes proportions en ce qui concerne les variables conjoncturelles (ou exogènes). Cet écart entre prévision et observation s'explique en partie par une hypothèse sous-jacente au modèle : quel sera le niveau moyen de rachat dans les mois à venir par rapport à aujourd'hui (ou période de référence à laquelle le modèle est construit) ? Le taux de rachat prévu sera ensuite ajusté en fonction de cette hypothèse de base. Le seuil d'affectation (0,5) de la réponse (cf section 2.2.4) pourrait aussi jouer sur ces mauvaises prévisions car la duplication des assurés crée un échantillon fortement déséquilibré avec 15571 rachats sur 991 010 observations, ce qui fait un taux de réponse nulle (non-rachat) de 98,43%. Néanmoins, cette hypothèse ne semble pas être à l'origine de la différence observée en 2007, pour la simple et bonne raison que la différence ne serait pas observée seulement en 2007.
L'aspect dynamique des rachats a également été traîté par des modèles fonctionnels (Ramsay & Silverman (2005) et Ramsay et al. (2009)) d'analyse de survie de type Cox et régression de Weibull (voir les excellents livres de Planchet & Thérond (2006) et Martinussen & Scheike (2006)) permettant de modéliser l'intensité de rachat à chaque

moment de la durée de vie du contrat, mais sans succès quant à la bonne prise en compte de l'influence du contexte extérieur. Pourtant cette approche permettait d'éviter de catégoriser la variable d'intérêt, à savoir l'ancienneté du contrat.

La conclusion à en tirer est que les prévisions de taux restent de qualité tant que les conditions économiques ne sont pas significativement différentes du passé, ce qui explique pourquoi l'usage de telles méthodes de prévisions ne s'est pas réellement popularisé dans la pratique actuarielle. La partie suivante illustre parfaitement le phénomène observé ici en 2007, et introduit une approche théorique pour le traitement de ce type de problématique. Cette théorie servira de point de départ dans la modélisation finale.

3.2 Impact de crises de corrélation des comportements

Les assureurs basent souvent leur modèle dynamique de taux de rachat sur une courbe déterministe en forme de S pour tenir compte de l'évolution des comportements de rachat en fonction des scénarios économiques (section 1.3.3). Cette courbe en S correspond au taux de rachat moyen exprimé en fonction de la différence entre deux taux, notée Δr. L'un de ces deux taux est le taux servi par l'assureur à ses assurés, tandis que l'autre peut valoir le taux du meilleur concurrent ou bien un taux d'intérêts (nous pourrions aussi imaginer que ce Δr représente une différence en termes de réputation...). L'idée courante est qu'un petit Δr ne provoque pas plus de rachats qu'à l'accoutumée, que le taux de rachat évolue de manière monotone et non-linéaire avec Δr, et que même si Δr est très grand certaines personnes resteront en portefeuille parce qu'elles ne prêtent pas spécialement attention à l'évolution des marchés. Le problème avec cette courbe en S est que nous n'avons jamais observé les comportements de rachat dans la situation extrême où Δr est très grand, ce qui implique que la construction d'un modèle stochastique s'appuie davantage sur des jugements d'expert que sur des données statistiques (qui n'existent tout simplement pas !).
Une manière simple d'introduire des effets stochastiques à cette courbe déterministe en S est de supposer une distribution gaussienne autour de la valeur du taux de rachat, mais cette section explique pourquoi il serait préférable d'utiliser une distribution bi-modale qui permette de prendre en compte le changement des corrélations entre comportements en scénarios extrêmes. Ces crises de corrélation (Biard et al. (2008), Loisel (2008)) suggèrent de ne pas utiliser l'approximation normale basée sur le théorème central limite (TCL). En effet ce théorème repose sur l'hypothèse de base que les décisions sont indépendantes les unes des autres, or ce ne serait vrai qu'avec la connaissance d'un facteur qui rende compte du niveau d'information des assurés, de la réputation de la compagnie et du secteur de l'Assurance. Ce facteur serait d'ailleurs clef pour comprendre la corrélation des risques de rachat avec d'autres risques tels que le risque de défaut ou de marché via la matrice de corrélation d'un modèle interne.
La crise du marché action et des produits dérivés a été suivie par une crise de corrélation : dans la plupart des cas, la corrélation grandit lors de scénarios défavorables. Il est probable qu'une situation extrême des taux d'intérêts conduise à des rachats massifs (tout du moins anormaux) suivant certaines déclarations politiques ou d'autres facteurs d'environnement (presse...). Par exemple l'une des premières phrases prononcée par les décideurs de pays développés suite au déclenchement de la crise fût : *Nous garantissons l'épargne des contribuables.* Cette attitude trahit leur crainte : ils anticipent des comportements extrêmes (loi binaire 0-1) plutôt qu'un comportement moyenné (gaussien). Nous nous appliquons dans la suite à développer un modèle simple qui tienne compte

de ces crises de corrélation : quand Δr grandit, la corrélation entre les décisions des assurés grandit et l'on passe d'une distribution en forme de cloche en régime classique à une distribution bimodale quand Δr devient grand. Nous présentons en premier lieu le modèle et son interprétation, puis des simulations et des formules analytiques de calcul de la distribution des taux de rachats sont fournies. Des résultats qualitatifs de l'impact de la corrélation sur la distribution du taux de rachat sont développés dans la dernière partie et en annexe C.1 via l'usage des ordres stochastiques, dans une optique de gestion de risque et de provisionnement basé sur un modèle interne.

3.2.1 Le modèle

Supposons que les assurés se comportent indépendamment avec un taux moyen de rachat $\mu(0)$ quand Δr vaut zéro, que le taux moyen de rachat vaut $1-\epsilon$ avec ϵ très petit quand Δr est très grand (disons 15%), et que la corrélation entre les décisions individuelles vaut $1-\eta$, avec η très petit. Le modèle suivant capture ces notions : soit I_k une variable aléatoire qui prend la valeur 1 si le k^{eme} assuré rachètent son contrat, 0 sinon. Supposons que

$$I_k = J_k I_0 + (1-J_k) I_k^{\perp},$$

où J_k correspond à l'indicatrice de l'événement "le k^{eme} assuré a un comportement moutonnier". La variable aléatoire J_k suit une loi de Bernoulli dont le paramètre p_0 est croissant en Δr, et I_0, $I_1^{\perp}$, $I_2^{\perp}$, ... sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.), dont le paramètre p est aussi croissant en Δr. Ainsi la probabilité de rachat croît avec Δr, et la corrélation (τ de Kendall ou ρ de Spearman) entre I_k et I_l (pour $k \neq l$) est égale à $P(J_k = 1 \mid \Delta r = x)$ quand $\Delta r = x$. Sans conditionner et donc en toute généralité, la corrélation entre I_k et I_l (pour $k \neq l$) vaut

$$\int_0^{+\infty} P(J_k = 1 \mid \Delta r = x)\, dF_{\Delta r}(x).$$

En effet sachant $\Delta r = x$, I_k et I_l (pour $k \neq l$) admettent une copule de Mardia (somme linéaire de la copule indépendante et de la borne supérieure de Fréchet) [*]. L'hypothèse gaussienne est plutôt juste quand $\Delta r = 0$ pour un portefeuille de 20 000 assurés. Nous allons voir avec des valeurs réalistes pour la courbe en S comment lorsque Δr augmente, la densité des taux de rachat évolue progressivement d'une forme en cloche vers une densité bimodale à partir d'un certain seuil $\Delta r = x_0$. McNeil et al. (2005) détaillent

∗. la copule d'I_k et I_l (pour $k \neq l$) n'est pas unique car leurs distributions ne sont pas continues.

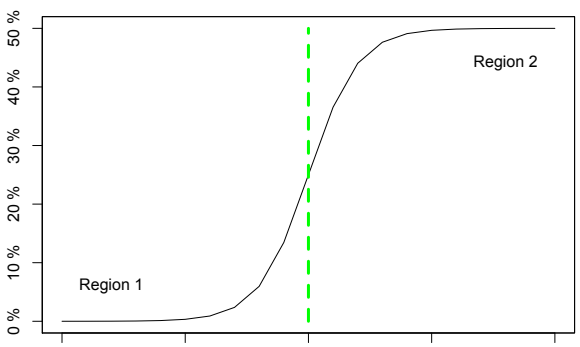


FIGURE 3.3 – Taux de rachat versus Δr.

précisément les problèmes de corrélation et leurs impacts sur la queue de la distribution de probabilité dans un contexte général.

3.2.2 Interprétation

La courbe en S du taux de rachat en fonction de Δr de la figure 3.3 signifie que moins le contrat est attractif et plus l'assuré a de chance de le racheter. La moyenne du taux de rachat est basse en régime économique de croisière (région 1, petit Δr sur la figure 3.3), et augmente significativement quand Δr croît. C'est la traduction d'une opportunité d'arbitrage que l'investisseur peut saisir : un contrat nouvellement acquis offre les mêmes garanties à un prix inférieur en cas de hausse des taux, ce qui mécaniquement améliore le rendement. Si à l'inverse les taux d'intérêt chutent, alors l'assureur peut choisir d'abaisser le taux crédité à l'assuré (suivant les modalités du contrat et pour des raisons financières, ou pour inciter les assurés à racheter).
Par conséquent la région 1 de la figure 3.3 correspond à la zone dans laquelle les décisions des assurés sont indépendantes (la corrélation tend vers 0), alors que la région 2 est celle des comportements corrélés (la corrélation tend vers 1). En fait la corrélation entre les comportements des assurés est quasi-nulle aussi longtemps que l'économie est en "bonne santé", le taux de rachat peut donc être modélisé par une loi normale dont la moyenne et l'écart-type sont ceux observés. C'est pourquoi la gaussienne visible en figure 3.4 est la distribution adaptée en région 1.
Inversement, la forte pente de la hausse du taux de rachat pour un certain niveau Δr en figure 3.3, suivie d'un plateau qui est le taux de rachat maximal atteignable (borne issue d'un jugement d'expert puisque jamais observée), reflète la détérioration des conditions économiques. Le point crucial consiste à réaliser que l'hypothèse d'indépendance est largement erronée dans un tel contexte : la corrélation entre les décisions des assurés fait changer la distribution du taux de rachat. C'est la conséquence de deux comportements extrêmement risqués dans lesquels presque tout le monde rachète ou quasiment personne ne rachète. La distribution la plus adaptée pour l'expliquer est la loi bimodale, illustrée en figure 3.4. La différence majeure avec le modèle gaussien est que la moyenne du taux résulte de deux pics de densité.
Remarquons qu'un comportement irrationnel des assurés peut également mener à des

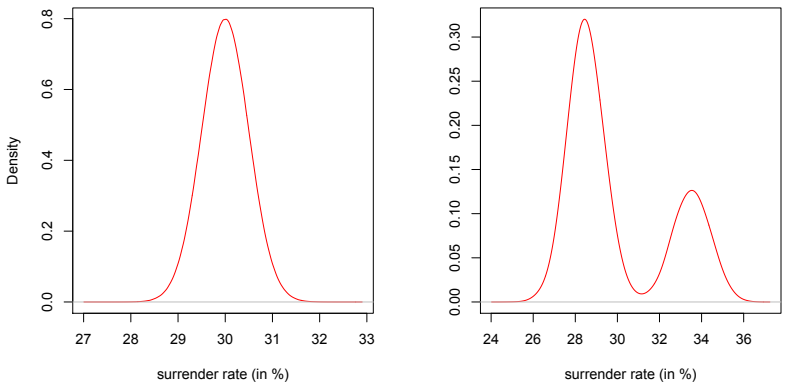


FIGURE 3.4 – Sur la gauche, la densité de la loi normale et sur la droite la densité bimodale (la moyenne vaut 30 dans les deux cas).

crises de corrélation même dans le cas où Δr est petit, ce qui (nous le verrons en section 3.2.4) est d'ailleurs la situation qui a le plus d'impact sur les besoins en capitaux ou augmentation des réserves de l'assureur. Un comportement irrationnel désigne ici un comportement atypique par rapport à l'expérience qu'a la compagnie, dû à des rumeurs ou des recommendations de journalistes ou brokers. D'un point de vue financier, un assuré adopte un comportement irrationnel s'il ne rachète pas son contrat bien qu'il soit gagnant dans cette opération. Mais ce comportement d'irrationalité (financier) n'est pas si rare que ça à cause des contraintes fiscales et de la complexité des contrats actuels d'Assurance-Vie, munis de garanties et d'options de plus en plus compliquées. Nous pouvons cependant remarquer que les agents semblent de plus en plus rationnels sur le marché américain (qui contient beaucoup de "variable annuities"), et que quelquepart l'incertitude concernant la rationalité future des assurés est capturé par notre modèle de crise de corrélation.

3.2.3 Distribution des taux de rachat

Approche combinatoire

Considérons un portefeuille de $n \geq 2$ assurés. Soit

$$N = \sum_{k=1}^{n} J_k$$

le nombre de personnes ayant un comportement moutonnier, et

$$M = \sum_{k=1}^{n} I_k$$

le nombre de personnes qui rachètent leur contrat. Rappelons que

$$I_k = J_k I_0 + (1 - J_k) I_k^{\perp},$$

où J_k correspond à l'indicatrice de l'événement "le k^{eme} assuré adopte un comportement moutonnier", et J_k a une distribution de Bernouilli de paramètre p_0, et où I_0, $I_1^{\perp}$, $I_2^{\perp}$, ... sont des variables aléatoires i.i.d. de paramètre p (et indépendantes des $(J_l)_{l\geq 1}$). Si le consensus général est de racheter ($I_0 = 1$), alors pour M valant un entier $k \in [0, n]$, le nombre N d'assurés "moutons" doit être inférieur ou égal à k, otherwise one would have $M \geq N > k$. Par le même raisonnement, si le comportement moutonnier consiste à ne pas racheter ($I_0 = 0$), alors pour M égal à un entier $k \in [0, n]$, le nombre N d'assurés "moutons" doit être inférieur ou égal à $n - k$, sinon nous aurions $M \leq n - N < n - (n - k) = k$. Nous obtenons à partir de la formula des probabilités totales que pour $0 \leq k \leq n$,

$$\begin{aligned} P(M = k) &= P(M = k \mid I_0 = 0) P(I_0 = 0) + P(M = k \mid I_0 = 1) P(I_0 = 1) \\ &= \sum_{i=0}^{k} P\left(M = k \mid I_0 = 1,\, N = i\right) P\left(I_0 = 1,\, N = i\right) \\ &+ \sum_{j=0}^{n-k} P\left(M = k \mid I_0 = 0,\, N = j\right) P\left(I_0 = 0,\, N = j\right). \end{aligned}$$

L'indépendance mutuelle entre les $(J_k)_{k\geq 1}$ et les $(I_l^{\perp})_{l\geq 1}$, avec $0 \leq k \leq n$ entraîne que

$$P(M=k) = p\sum_{i=0}^{k} a_{i,k} + (1-p)\sum_{j=0}^{n-k} b_{j,k},$$

avec pour $0 \leq i \leq k$,

$$a_{i,k} = C_n^i p_0^i (1-p_0)^{n-i} C_{n-i}^{k-i} p^{k-i} (1-p)^{n-k},$$

et pour $0 \leq j \leq n-k$

$$b_{j,k} = C_n^j p_0^j (1-p_0)^{n-j} C_{n-j}^{k} p^{k} (1-p)^{n-j-k}.$$

Remarquons que pour k fixé, les $a_{i,k}$, $0 \leq i \leq k$ et les $b_{j,k}$, $0 \leq j \leq n-k$ peuvent être calculés grâce aux formules récursives suivantes : pour $0 \leq i \leq k$, nous avons

$$\frac{a_{i+1,k}}{a_{i,k}} = \frac{C_n^{i+1}}{C_n^i}\frac{p_0}{p\,(1-p_0)}\frac{C_{n-i-1}^{k-i-1}}{C_{n-i}^{k-i}} = \frac{k-i}{i+1}\frac{p_0}{p\,(1-p_0)}$$

et pour $0 \leq j \leq n-k$, nous avons

$$\frac{b_{j+1,k}}{b_{j,k}} = \frac{C_n^{j+1}}{C_n^j}\frac{p_0}{(1-p)\,(1-p_0)}\frac{C_{n-j-1}^{k}}{C_{n-j}^{k}} = \frac{n-j-k}{j+1}\frac{p_0}{(1-p)\,(1-p_0)}.$$

Notons qu'il est préférable de commencer avec a_{i_0} and b_{j_0} tels que a_{i_0} and b_{j_0} soient assez grands dans le but de minimiser les erreurs d'arrondis lors du calcul de

$$a_0 = b_0 = (1-p_0)^n\, C_n^k p^k (1-p)^{n-k}$$

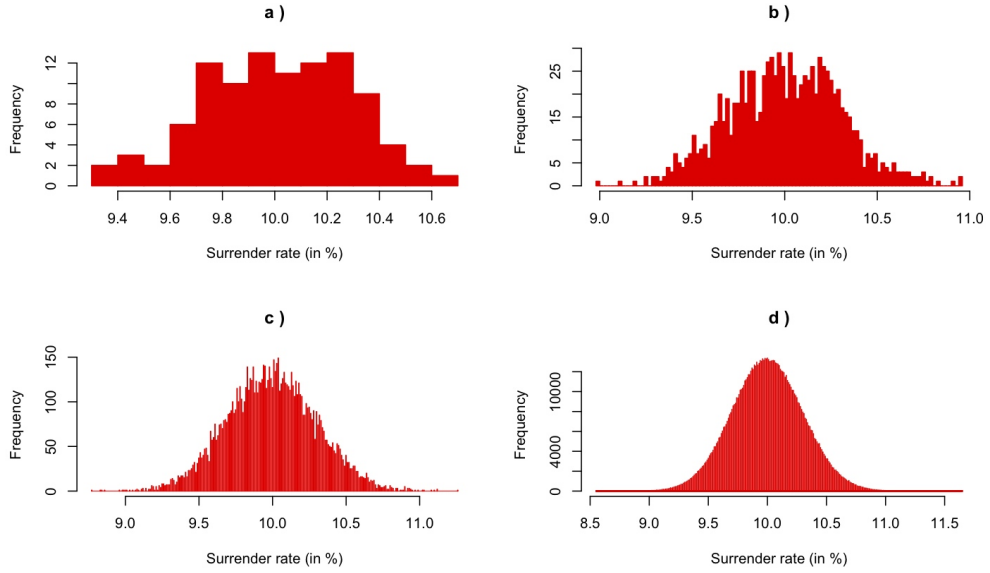
qui sont en général assez petits. Viquerat (2010) propose des algorithmes efficaces (et leur précision) pour effectuer ce type de calcul.

Approche par simulation

En pratique, l'utilisation de simulations est courante pour l'évaluation du risque de rachat, parmi les nombreux types de risque d'un modèle interne complexe. Le nombre de simulations est clef pour obtenir une approximation précise de la distribution du taux de rachat, quelque soit le contexte socio-économique. Le nombre d'assurés en portefeuille a son importance car il permet de diminuer la dispersion des valeurs de taux de rachat, bien que cela n'affecte pas vraiment la forme de la distribution. La figure 3.5 confirme ces remarques, nous prendrons donc dans la suite un nombre de simulations égal à 1000 000 et un nombre d'assurés de 10 000. Il va sans dire de l'effet capital de la probabilité individuelle de rachat p, qui joue directement sur la moyenne de la distribution et engendre un profil plus risqué.

Concentrons nous maintenant sur l'effet de la corrélation, le coeur de ce chapitre. Le paramétre de corrélation p_0 (probabilité de suivre le consensus collectif) joue également un rôle crucial : la hausse de p_0 remodèle la forme de la distribution du taux de rachat. Une crise économique provoque naturellement l'augmentation simultanée de la probabilité de rachat et de la corrélation, ce qui est une très mauvaise nouvelle pour l'assureur qui doit faire face à une situation dans laquelle les modes s'équilibrent (risque élevé dans les deux cas) et s'écartent. La figure 3.6 illustre cette déformation.

FIGURE 3.5 – Effet du nombre de simulations sur la distribution du taux de rachat : a) 100, b) 1 000, c) 10 000 and d) 1 000 000. Pas de comportement moutonnier, probabilité individuelle de rachat égale à 10%, 10 000 assurés en portefeuille.



Pour un certain Δr (et donc pour un p_0 donné en théorie), plus le paramètre de corrélation p_0 est grand et plus la forme de la densité de rachat devient bimodale. Tout l'intérêt de l'assureur se porte sur la quantification de la différence entre ces distributions en termes de risque de comportement. A cette fin, certaines mesures de risque dont la

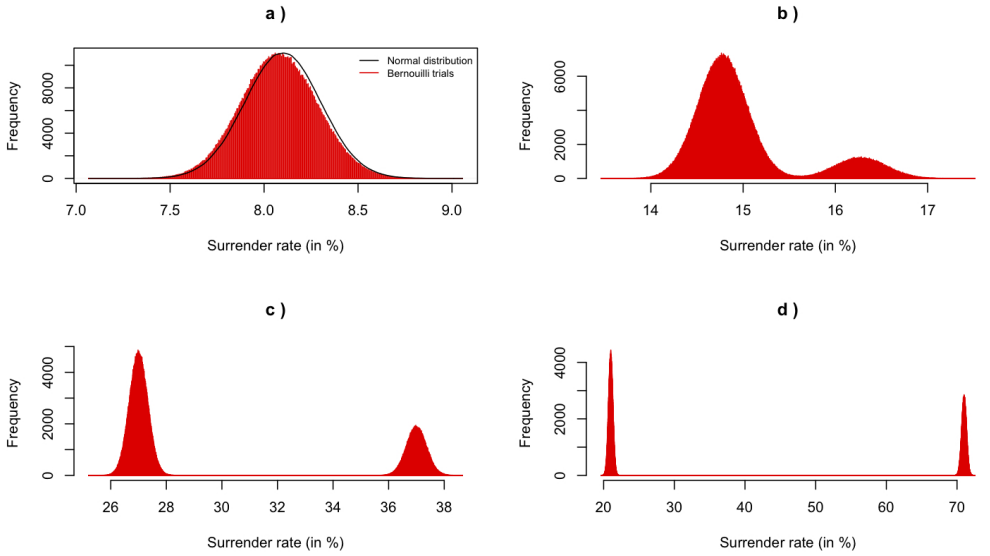
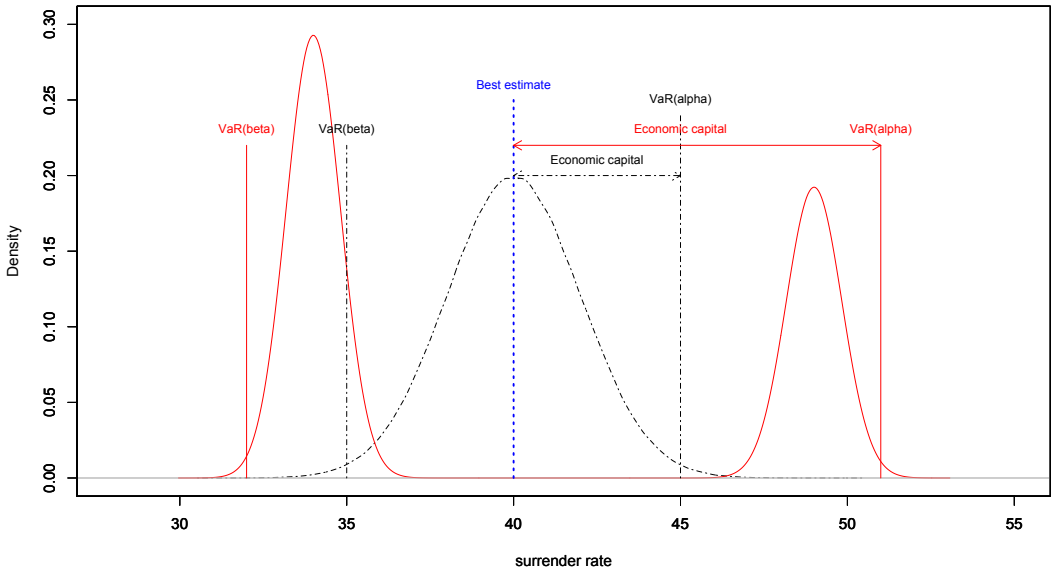


FIGURE 3.6 – Evolution d'une distribution gaussienne de taux de rachat à une bimodale. De haut en bas et de gauche à droite, les probabilités de rachat et de comportement moutonnier valent : a) 8% et 0%, b) 15% et 1.5%, c) 30% et 10% et d) 42% et 50%. 1 000 000 simulations et 10 000 assurés.

FIGURE 3.7 – Densité du taux de rachat pour des comportements indépendants (en noir et pointillé) et pour des comportements corrélés (en rouge et trait plein) : le capital économique associé (qui vaut la différence entre la VaR_α et le "best estimate") s'accroît avec la corrélation.



Value-at-Risk (ou *VaR*) peuvent servir d'indicateurs de cet écart. La *VaR* est définie pour une variable aléatoire X et un seuil α par :

$$VaR_\alpha(X) = \inf\{x \in X,\ F_X(x) \geq \alpha\}.$$

La variable aléatoire X est le taux de rachat dans notre application, ce qui signifie que l'assureur s'attend à subir un taux de rachat inférieur à *VaR* avec $\alpha\%$ de confiance. En Assurance-Vie, nous posons en général $\alpha = 99.5\%$. Parfois la situation antagoniste (chute du taux de rachat) préoccupe également l'assureur car son exposition peut devenir trop grande par rapport à ses contraintes de capitaux (par exemple les garanties décès sont évaluées avec certaines prévisions d'exposition au risque basées sur les rachats du passé), ou parce que les taux d'intérêts lui sont défavorables (sur des produits à taux garantis). Dans ce cas, nous adaptons le raisonnement en considérant les VaR_α (côté droit, risque de rachats massifs) et VaR_β (côté gauche, très peu de rachat comparé aux prévisions) illustrées en figure 3.7.

3.2.4 Ecarts de VaR et taille du portefeuille

Dans une perspective Solvabilité II, nous nous focalisons sur une analyse détaillée des écarts de VaR à 99.5 %. La figure 3.8 récapitule l'effet de l'accroissement de la corrélation entre décisions des assurés sur la VaR. Pour une probabilité individuelle de rachat donnée, disons d'1%, une corrélation passant de 0 à 1% augmente la $VaR_{99.5\%}$ de 30 à 50%. Nous pouvons notamment remarquer que les grands écarts **positifs** de VaR sont concentrés sur de petites valeurs de corrélation lorsque nous considérons une faible propension au rachat. Ce résultat remarquable nous suggère de définir des classes de risque en termes de *sensibilité* (par rapport à la corrélation) :

- *hyper sensible* (zone rouge dans la figure 3.8) : $p \in]0, 0.05]$ et $p_0 \in]0, 0.1]$;
- *sensible* (zone orange) : $p \in]0, 0.05]$ et $p_0 \in [0.1, 0.4]$, ou $p \in]0.05, 0.2]$ et $p_0 \in]0, 0.3]$;

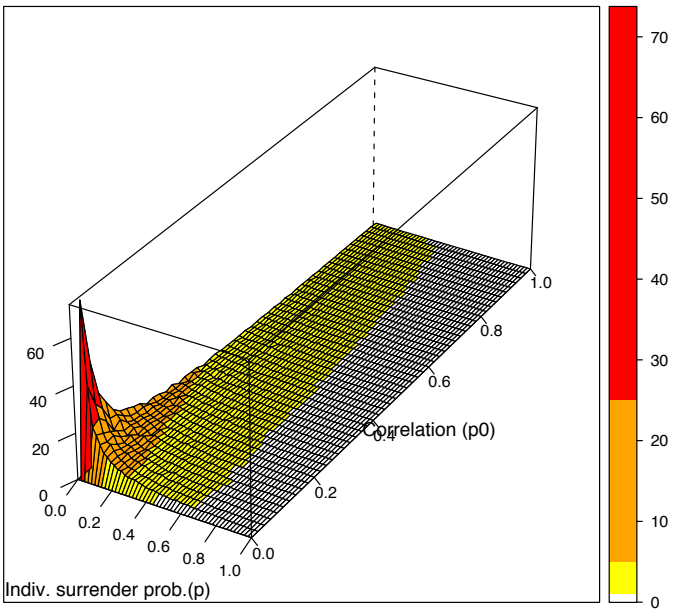


FIGURE 3.8 – Ecart relatif (en %) des VaR versus p_0 et p.

– *peu sensible* (zone jaune) : autres situations.

Dans la configuration *hyper sensible*, la $VaR_{99.5\%}$ peut augmenter jusqu'à 70% ! Dans la configuration *sensible*, l'assureur peut voir sa $VaR_{99.5\%}$ augmenter de 5 to 25 %, ce qui est moins risqué mais reste très préoccupant. Enfin, la configuration *peu sensible* est une zone dans laquelle l'assureur peut être serein car il semble avoir déjà assez provisionné (la VaR est assez grande). Ces observations montrent que que la situation la plus dangereuse en termes d'écart de provisionnement pour l'assureur correspond à l'apparition de la corrélation dans des scénarios où la probabilité moyenne de rachat est très faible.

La taille de la compagnie pourrait également être un facteur-clef : pour tester son impact, nous avons simulé les écarts de capital économique (EC) lié à la VaR pour une probabilité annuelle moyenne de rachat réaliste de 8,08 % (BE pour *best-estimate* dans le tableau 3.1) et une corrélation passant de 0 % à 50 %. Etudier les écarts de capitaux économiques revient à étudier les écarts de VaR (voir graphe 3.7). Certaines compagnies d'assurance pourraient penser que leur taille leur évite des scénarios catastrophiques

Taille portefeuille	BE	EC ($VaR^{Normale}_{99.5\%}$)	Corrélation	EC ($VaR^{Bimodale}_{99.5\%}$)	ΔEC
Petite :			$p_0 = 0.05$	6.26%	4.5%
5000	8.08%	1.76%	$p_0 = 0.2$	20.42%	18.66%
assurés	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	48.54%	46.78%
Moyenne :			$p_0 = 0.05$	5.1%	4.59%
50 000	8.08%	0.51%	$p_0 = 0.2$	19.01%	18.5%
assurés	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	46.63%	46.12%
Grande :			$p_0 = 0.05$	4.73%	4.59%
500 000	8.08%	0.1426%	$p_0 = 0.2$	18.56%	18.41%
assurés	($p_0 = 0$)	($p_0 = 0$)	$p_0 = 0.5$	46.16%	46.01%

TABLE 3.1 – Impact de la taille du portefeuille sur la VaR (100 000 simulations).

grâce à l'effet de la mutualisation. L'analyse du tableau 3.1 démontre que le nombre d'assurés en portefeuille n'a pas de forte influence sur le calcul des marges de risque. En effet, la différence de capital économique ΔEC nécessaire baisse de manière dérisoire, même en passant de 5 000 à 500 000 assurés !

3.3 Conclusion

Nous avons montré dans cette partie que l'impact du choix de la distribution des rachats est majeur tant pour les calculs de besoin en capital économique que pour les prévisions de taux de rachat. Cette distribution résulte de deux points de vue opposés sur la modélisation des comportements : de comportements supposés indépendants, nous évoluons vers une modélisation de comportements de rachat corrélés qui selon nous reflète davantage la réalité, comme l'illustre le graphique 3.1 (un exemple numérique concret est étudié dans Loisel & Milhaud (2011)). Nous pourrions aussi investiguer l'impact rétroactif des rachats massifs sur les taux d'intérêt et d'inflation pour éventuellement détecter un cercle vicieux, ou encore considérer les rachats partiels provenant d'options contenues dans les contrats qui permettent de basculer une partie de son épargne vers un autre type de support (ex : transférer de l'UC vers un fonds Euro, Loisel (2010)). Cette dernière information serait idéale mais n'est concrètement jamais accessible.
En résumé, provisionner assez d'argent dans le but de couvrir le risque de corrélation des comportements est très important dans la prévention des besoins en capitaux. En effet il pourrait découler de la sous-estimation de ce risque des appels de marge aux actionnaires, ce qui serait très néfaste pour la compagnie. De plus, le caractère non-diversifiable de certains risques fait que la taille du portefeuille ne permet pas de réduire l'impact des crises de corrélation sur les quantités considérées.

L'approche théorique *supra* nous conduit naturellement à l'extension développée dans le chapitre suivant : les mélanges de régressions logistiques, permettant de capter à la fois les phénomènes de corrélation et l'hétérogénéité des comportements entre cohortes.

Chapitre 4

Mélange de régressions logistiques

Le risque de rachat est très compliqué à étudier et à modéliser car c'est un risque de comportement humain qui dépend de nombreux facteurs : les caractéristiques individuelles, les désirs et besoins personnels, les options du contrat, son ancienneté, le contexte économique et financier (problèmes de liquidité), les aspects socio-culturels (ex : comparaison expérience Japon / Etats-Unis), mais aussi les décisions du régulateur. Les deux chapitres précédents nous ont permis de pointer du doigt les trois problématiques majeures : la dimension des données, les problèmes de corrélation entre comportements et l'hétérogénéité des décisions face à un environnement évolutif. Les rachats s'expliquent à la fois par des caractéristiques idiosynchratiques mais aussi par un ensemble de facteurs exogènes dont certains sont très difficilement quantifiables (réputation), voire inaccessible (politique de vente future). Une méthodologie quant à la réduction de la dimension des données se détache par l'utilisation des algorithmes CART qui ont l'avantage de ne pas supposer d'hypothèses fortes sous-jacentes aux données (relation linéaire, log-linéaire,...), et qui ont démontré leur robustesse. Les quelques variables sélectionnées (nous nous limitons à deux ou trois variables en général car un modèle surdimensionné donne souvent de mauvaises prévisions) présentant le plus fort impact sur l'événement de rachat seront ensuite introduites dans des modèles dynamiques. Nous avons également compris que l'association de facteurs exogènes et endogènes dans une même et unique équation de régression ne permet pas de capter les bons effets, le défi étant donc de trouver une manière fonctionnelle de considérer ces effets différemment. Ce chapitre propose donc une extension des premières modélisations évoquées, dans le sens où nous "mixons" nos idées pour parvenir à nos fins.

Les modèles mélange sont une technique populaire pour modéliser l'hétérogénéité non-observable de données ou pour approximer une distribution générale de manière semi-paramétrique. Ils sont utilisés dans de nombreux domaines d'application tels que l'économie, l'astronomie, la biologie, la médecine. Historiquement, les mélanges ont été introduit pour la première fois il y a plus de cent ans par Pearson (1894). Karl Pearson utilise un mélange de lois normales dans le cadre de la modélisation de la longueur du corps des crabes. La modélisation de données asymétriques est aussi réalisable via des transformations de données, notamment la transformation en *log* (Box & Cox (1964)). Il est souvent difficile de faire un distinguo entre des données présentant une certaine asymétrie et des données provenant d'un mélange (McLachlan & Peel (2000), bas p.15), bien que dans notre cas l'asymétrie est telle qu'il ne fait aucun doute qu'une simple transformation ne pourrait pas modéliser toute l'hétérogénéité présente dans nos

données. Le cas le plus classique de la modélisation par mélange concerne les mélanges de lois normales, pour lesquels un grand nombre de résultats existe. Les modèles mélange sont aussi régulièrement utilisés dans des problèmatiques de classification puisqu'ils assignent un groupe donné à une observation, .

La première partie développe l'aspect théorique des modèles mélange, les points importants à aborder lors de leur utilisation et les pièges à éviter. En ce qui concerne l'application, nous verrons ensuite les outils pratiques de présentation des résultats de la modélisation à travers un cas pratique (toujours basé sur les contrats mixtes en Espagne).

4.1 Formalisation de la théorie

La modélisation par mélange renvoie aux problèmes usuels suivants : identifiabilité, estimation des paramètres, propriétés de l'estimateur du maximum de vraisemblance, évaluation du nombre de composantes du mélange, application de la théorie asymptotique pour fournir une base de solutions à certains problèmes, critères de sélection et de performance du modèle. L'estimation des paramètres d'un mélange est un des axes de recherche ayant attiré le plus de chercheurs car de nombreuses questions subsistent encore aujourd'hui ; parmi lesquelles les valeurs initiales de l'algorithme d'optimisation qui maximise la vraisemblance, les critères d'arrêt de cet algorithme et les propriétés de la fonction de vraisemblance (convexité, bornitude). Nous allons dans cette partie tenter de résumer l'ensemble de ces problématiques afin de donner au lecteur une base théorique qui lui permette d'appréhender ce type de modélisation.

4.1.1 Généralités

Nous formalisons l'approche par mélange dans le cadre d'un mélange discret car elle est bien plus intuitive. Néanmoins, toutes les notions développées ci-dessous peuvent être adaptées en cas continu, ce qui veut dire que la distribution mélangeante est continue (nous verrons qu'elle est multinomiale dans le cas discret). De plus, nous nous placerons dans un contexte de mélange discret pour résoudre notre problématique opérationnelle.

Soit $Y = (Y_1^T, ..., Y_n^T)^T$ un échantillon aléatoire. Chaque enregistrement Y_j de cet échantillon contient p mesures, d'où un vecteur aléatoire p-dimensionnel ($p = 1$ pour nous car la réponse est univariée). Dans le contexte des mélanges et par la formule des probabilités totales, il vient

$$f(y_j) = \sum_{i=1}^{g} \pi_i f_i(y_j), \tag{4.1}$$

où $f(y_j)$ est la densité de Y_j dans $\mathbb{R}^p$, π_i est la proportion (poids) **à-priori** de la i^{eme} composante du mélange, $f_i(y_j)$ est la densité de la i^{eme} composante du mélange, avec la contrainte $\sum_i \pi_i = 1$. La matrice Y des observations est de taille $n \times p$. On dit que $f(y_j)$ est la densité d'un mélange fini à g composantes, et on note $F(y_j)$ la distribution du mélange. Chaque individu est donc censé provenir d'un des groupes composant le mélange.

En général, g est fini mais inconnu et doit donc être estimé inférentiellement à partir des données. Les probabilités d'appartenance à tel ou tel groupe doivent être estimées en même temps, de même que les densités $f_i(.)$. Pour comprendre l'interprétation de la modélisation par mélange, une bonne méthode consiste à essayer de le générer. Pour

simuler la variable Y_j, nous définissons la variable Z_j d'appartenance à une composante par

$$Z_j = \begin{cases} 1 & \text{avec probabilité } \pi_1 \text{ si l'individu j appartient au groupe 1,} \\ 2 & \text{avec probabilité } \pi_2 \text{ si l'individu j appartient au groupe 2,} \\ ... & \\ g & \text{avec probabilité } \pi_g \text{ si l'individu j appartient au groupe g,} \end{cases}$$

et la densité conditionnelle de Y_j est donnée par $f_{Y_j|Z_j=i}(y_j) = f_i(y_j)$. Nous pouvons donc voir Z_j comme le vecteur aléatoire $Z_j = (Z_{j1}, Z_{j2}, ..., Z_{jg})^T$ où

$$Z_{ij} = (Z_j)_i = \begin{cases} 1 & \text{si la composante d'appartenance de } Y_j \text{ dans le mélange est la } i^{eme}, \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, Z_j suit une loi multinomiale et l'on note $Z_j \sim Mult_g(1, \pi)$ avec $\pi = (\pi_1, ..., \pi_g)^T$. Nous avons donc

$$P(Z_j = z_j) = \pi_1^{z_{1j}} ... \pi_g^{z_{gj}}.$$

Les mélanges peuvent être vus comme une alternative entre un modèle complètement paramétrique et un modèle non-paramétrique. Dans le cas non-paramétrique, nous retrouvons l'estimateur à noyau de la densité en prenant $g = n$ composantes (où n est le nombre d'observations), des poids tous égaux $\pi = 1/n$ et une densité $f_i(y_j) = \frac{1}{h} k(\frac{y_j - y_i}{h})$ où $k()$ est une densité. A l'inverse, si l'on fixe $g = 1$ composante, alors le modèle devient complètement paramétrique. Nous nous intéressons dans la suite aux cas où $g \in [1; n]$.

Nous l'avons dit en introduction : la multimodalité des données peut ne pas provenir d'un mélange. Il est possible de détecter ceci par l'usage du test du ratio de vraisemblance, mais la difficulté vient du fait que nous ne connaissons pas la distribution de la statistique de test sous l'hypothèse nulle dans ce cadre-là. Nous utilisons alors une approche de reéchantillonage qui permet d'obtenir une *p-valeur* de test sans connaître cette statistique (McLachlan & Peel (2000), p.75). La clef pour l'estimation des paramètres d'un mélange est de reformaliser le problème de données incomplètes sous forme d'un problème aux données complètes : en effet, nous ne connaissons pas le groupe d'appartenance de chaque observation dans la réalité, mais l'introduction de la variable Z_j va nous permettre de mener directement l'estimation par maximum de vraisemblance par l'algorithme espérance-maximisation (EM). Dans un contexte bayésien (qui ne sera pas le notre), cette vision du problème permet d'estimer les paramètres par des méthodes MCMC (Monte Carlo Markov Chain).

En résumé, nous observons $y = (y_1, ..., y_n)$, réalisations de $Y = (Y_1, ..., Y_n)$ issues de la même densité mélange donnée par (4.1). Ces observations sont i.i.d. et nous avons

$$Y_1, ..., Y_n \sim F = F(Y_j).$$

Les données complètes, notées y_c, s'exprimeraient donc comme $y_c = \begin{pmatrix} (y_1, z_1) \\ (y_2, z_2) \\ ... \\ (y_n, z_n) \end{pmatrix}$. Grâce à la formule de Bayes, nous pouvons calculer la probabilité **à-posteriori** d'appartenir

à telle ou telle composante du mélange :

$$\begin{aligned}\tau_i(y_j) &= P(y_j \in \text{ composante i } |y_j) \\ &= P(Z_{ij}=1|y_j) \\ &= \frac{P(Z_{ij}=1 \cap y_j)}{P(y_j)} = \frac{P(y_j|Z_{ij}=1)P(Z_{ij}=1)}{P(y_j)} \\ &= \pi_i \frac{f_i(y_j)}{f(y_j)}, \end{aligned} \tag{4.2}$$

pour $i = 1, .., g$ et $j = 1, ..., n$.
En pratique, nous estimons les π_i par leur moyenne empirique, i.e. $\hat{\pi}_i = \sum_{j=1}^{n} z_{ij}/n$; et les paramètres des composantes du mélange par les données qui y appartiennent.
Une formulation paramétrique d'un modèle mélange peut s'écrire de la manière suivante

$$f(y_j) = f(y_j; \Psi) = \sum_{i=1}^{g} \pi_i f_i(y_j; \theta_i), \tag{4.3}$$

avec $\Psi = (\pi_1, ..., \pi_{g-1}, \xi^T)^T$ et $\xi^T = (\theta_1^T, ..., \theta_g^T)$. Nous noterons Ω l'espace des paramètres de Ψ. Faisons l'hypothèse que les composantes appartiennent à la même famille paramétrique, nous considérons une distribution mélangeante discrète $H(\theta)$ définie par $H(\theta) = P(\theta = \theta_i) = \pi_i$ pour i=1,...,g. Alors le modèle mélange se reécrit comme

$$f(y_j; H) = \int f(y_j; \theta) dH(\theta).$$

Pour généraliser, nous pouvons considérer une mesure de probabilité plus générale pour H (une loi continue par exemple).

Il existe dans la littérature plusieurs techniques d'estimation de la distribution mélange : la méthode graphique, la méthode des moments, la méthode des distances minimum, l'approche bayésienne et le maximum de vraisemblance. Cette grande variété est dûe au fait que nous n'avons pas de formules explicites pour les estimateurs, qui sont calculés itérativement par divers algorithmes. La taille n de l'échantillon doit être relativement grande pour garantir les propriétés asymptotiques des mélanges.

4.1.2 Identifiabilité

L'estimation de ψ sur la base des observations y_j n'a de sens que si Ψ est identifiable. La définition de l'identifiabilité dans le cadre des mélanges diffère un peu du cas classique dans la mesure où il y a la notion supplémentaire de composantes. Intuitivement, un modèle est identifiable si des valeurs distinctes de Ψ déterminent des membres distincts de la famille paramétrique associée à Ψ (il ne peut pas y avoir deux paramètres différents qui donnent le même modèle à l'arrivée). Formellement, $f(y_j; \Psi)$ est une famille paramétrique de densité identifiable dans le cadre classique si

$$f(y_j; \Psi) = f(y_j; \Psi') \Leftrightarrow \Psi = \Psi'$$

Pour les mélanges, la seule définition ne suffit pas car on peut avoir une classe de mélange identifiable sans pour autant avoir identifiabilité sur Ψ. Il suffit pour comprendre cela de permuter les composantes d'appartenance (les "labels"), ce qui ne change pas la densité globale du mélange mais qui rend Ψ non-identifiable pour des densités composantes appartenant à la même famille paramétrique. Il faut donc ajouter une contrainte

supplémentaire. Soit $f(y_j;\Psi)=\sum_{i=1}^{g}\pi_i f_i(y_j;\theta_i)$ et $f(y_j;\Psi^\star)=\sum_{i=1}^{g}\pi_i^\star f_i(y_j;\theta_i^\star)$ deux membres d'une famille paramétrique de mélange. Cette classe de mélanges finis est dite **identifiable** pour $\Psi\in\Omega$ si

$$f(y_j;\Psi)=f(y_j;\Psi^\star)$$
$$\Updownarrow$$
$g=g^\star$ et on peut permuter les indicatrices de composantes d'appartenance pour que $\pi_i=\pi_i^\star$ et $f_i(y_j;\theta_i)=f_i(y_j;\theta_i^\star)$.

En général, nous ajoutons des contraintes pour palier au manque d'identifiabilité dû aux permutations possibles entre composantes d'appartenance. Un détail important est à ajouter : le manque d'identifiabilité est un problème important en analyse bayésienne des mélanges lors de la simulation à-posteriori de l'appartenance à un groupe donné, mais n'est pas préoccupant dans le cadre de l'estimation par maximum de vraisemblance.

En dehors de l'identifiabilité, un autre problème à ne pas confondre est celui de l'identification : est-ce facile de savoir à quelle composante appartient une observation donnée ? La réponse dépend évidemment de la répartition des données. Une multimodalité prononcée sera moins problématique que des données faiblement asymétriques.

4.1.3 Algorithme espérance-maximisation (EM)

Cet algorithme (aussi connu sous le nom d'échantillonneur de Gibbs dans le cadre bayésien) ultra-connu offre des propriétés très intéressantes pour l'optimisation de fonction de vraisemblance complexe, sur un problème aux données manquantes. Ces propriétés ont été démontré dans un article célèbre de Dempster et al. (1977), qui a permis avec la révolution informatique l'explosion de l'usage de ce type de modèle, qui jusque là demandait de complexes et fastidieux calculs pour maximiser la vraisemblance. Nous donnons ici la version originelle de cet algorithme et son idée, sachant qu'une multitude de développement ont depuis été proposés pour traiter des problématiques particulières (convergence vers le maximum global, dimension des données, y_j manquantes, ...). Le principe de base de cet algorithme est de transformer le problème aux données manquantes en problème aux données complètes $Y_c=(Y^T,Z^T)^T$ où les $Z_j\sim Mult_g(1,\pi)$ et sont i.i.d. La log-vraisemblance des données complètes pour une observation j vaut $f(y_{jc};\Psi)=\Pi_{i=1}^{g}[\pi_i f_i(y_j;\theta_i)]^{z_{ij}}$, d'où la log-vraisemblance des donnés complètes sur l'échantillon entier $\log L_c(\Psi)=\log(\Pi_{j=1}^{n}f(y_{jc};\Psi))$ qui donne après développement :

$$\log L_c(\Psi)=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}[\log\pi_i+\log f_i(y_j;\theta_i)]. \tag{4.4}$$

Etape espérance Traitement de la donnée z_j non observable par l'espérance conditionnelle de $\log L_c(\Psi)$, sachant ce que nous observons (y). Soit $\Psi(0)$ la valeur initiale de Ψ. Nous calculons

$$Q(\Psi,\Psi^{(0)})=\mathbb{E}_{\Psi^{(0)}}[\log L_c(\Psi)\mid y],$$

or $\log L_c(\Psi)$ est linéaire en z_{ij}, donc l'étape espérance requiert uniquement le calcul de $\mathbb{E}[Z_{ij}\mid y]$. Nous avons

$$\mathbb{E}_{\Psi^{(k)}}[Z_{ij}\mid y]=P_{\Psi^{(k)}}(Z_{ij}=1\mid y)=\tau_i(y_j;\Psi^{(k)}), \tag{4.5}$$

avec τ_i la probabilité à-posteriori à l'étape k. En injectant (4.5) dans (4.4), nous pouvons calculer à l'étape $(k+1)$ l'expression de Q :

$$\begin{aligned} Q(\Psi, \Psi^{(k)}) &= \sum_{i=1}^{g}\sum_{j=1}^{n} \mathbb{E}[Z_{ij} \mid y] \times [\log \pi_i + \log f_i(y_j; \theta_i)], \\ Q(\Psi, \Psi^{(k)}) &= \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)}) \times [\log \pi_i + \log f_i(y_j; \theta_i)], \end{aligned} \tag{4.6}$$

où nous avons avec (4.2),

$$\tau_i(y_j; \Psi^{(k)}) = \pi_i^{(k)} \frac{f_i(y_j; \theta_i^{(k)})}{f(y_j; \Psi^{(k)})} = \pi_i^{(k)} \frac{f_i(y_j; \theta_i^{(k)})}{\sum_{h=1}^{g} \pi_h^{(k)} f_h(y_j; \theta_h^{(k)})}.$$

Etape maximisation A l'étape $k+1$, nous voulons maximiser globalement $Q(\Psi, \Psi^{(k)})$ par rapport à Ψ sur Ω, pour donner une estimation $\Psi^{(k+1)}$. Dans le cas des mélanges finis, nous estimons séparément les proportions et les densités composantes. Il vient :

$$\begin{cases} \pi_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)}) : \text{moyenne empirique des probabilités à-posteriori,} \\ \xi^{(k+1)} \text{ est obtenu en résolvant } \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)}) \dfrac{\delta \log f_i(y_j; \theta_i)}{\delta \xi}. \end{cases}$$

La propriété de d'accroissement monotone de la vraisemblance des données $Q(\Psi, \Psi^{(k)})$ à chaque étape garantit la convergence de la vraisemblance vers une valeur stationnaire (maximum local ou global). Nous répétons successivement ces étapes jusqu'à ce que le critère d'arrêt de l'algorithme soit satisfait, en général $L_c(\Psi^{(k+1)}) - L_c(\Psi^{(k)})$ plus petit qu'un certain seuil. Toutefois cette procédure d'arrêt n'est pas toujours satisfaisante, c'est pourquoi Lindstrom & Bates (1988) et Böhning et al. (1994) propose une amélioration basée sur le critère d'accélération de Aitken. Pour détecter la pertinence des estimations trouvées, il n'existe pas de méthode prédéfinie : la solution consiste à regarder à la fois la valeur de la vraisemblance, les valeurs des π_i estimées et les matrices de covariance (voir les exemples p.100 de McLachlan & Peel (2000)).
La vitesse de convergence de l'algorithme EM dépend de la proportion d'information manquante sur Ψ du fait que l'on observe seulement les réalisations de Y au lieu d'observer conjointement Y et Z ; plus cette proportion est grande et plus l'algorithme est lent. Nous ne discutons pas ici des variantes de l'EM qui permettent de contourner les problèmes de valeurs initiales de l'algorithme, mais le lecteur intéressé pourra trouver son bonheur dans McLachlan & Peel (2000).

4.1.4 Evaluation du nombre de composantes

L'évaluation du bon nombre de composantes d'un modèle mélange a toujours été difficile et le problème n'est pas encore vraiment résolu. Les mélanges ont principalement deux fonctions : fournir une classification basée sur une modélisation, et définir une méthode semi-paramétrique permettant de modéliser des formes de distribution inconnues comme une alternative à la méthode des noyaux. Dans ces deux approches, comment choisir g ?
Nous avons pu constater la séparation du problème de l'évaluation de g et celui de l'estimation des paramètres, dans le sens où l'on fixe d'abord g avant de lancer l'estimation. Nous faisons cela pour plusieurs valeurs de g. L'usage commun pour trouver g est :

– de considérer des critères de sélection tels que le critère d'information de Akaike ou le Bayesian Information Criterion (respectivement AIC et BIC),
– de se servir du test du ration de vraisemblance (LRT),

mais il existe aussi des méthodes non-paramétriques, la méthode des moments, l'approche basée sur le Kurtosis de la distribution...Les références à toutes ces techniques sont disponibles dans l'ouvrage de McLachlan & Peel (2000). Nous ne trouvons pas utile de discuter davantage des critères AIC et BIC car ils sont très connus et un large panel de papiers traîte de leur utilisation.

En revanche nous souhaitons préciser en quoi consiste le LRT dans le cadre des mélanges, sans pour autant entrer dans trop de détails. Ce test a pour but de trouver la plus petite valeur convenable de g, avec comme hypothèses nulle et alternative :

$$[\quad H_0 : g = g_0 \quad \text{contre} \quad g = g_1 \quad], \text{ avec } g_1 > g_0.$$

En pratique nous prenons $g_1 = g_0 + 1$ et nous continuons d'ajouter des composantes tant que l'accroissement de la valeur de la vraisemblance est substantiel. Soient $\hat{\Psi}_1$ l'estimateur par maximum de vraisemblance (MLE) de Ψ sous H_1, et $\hat{\Psi}_0$ le MLE sous H_0. Nous notons

$$-2\log\lambda = 2[\log L(\hat{\Psi}_1) - \log(\hat{\Psi}_0)] = 2\log\frac{\hat{\Psi}_1}{\hat{\Psi}_0}.$$

Si λ est suffisamment petite, ou si $-2\log\lambda$ est suffisamment grand, il parait logique de pouvoir rejeter H_0. Malheureusement ici, nous ne connaissons pas dans ce cas la distribution nulle de $-2\log\lambda$ dans le cas général, car les conditions de régularité nécessaires (Cramér (1946)) aux résultats asymptotiques du MLE ne sont pas satisfaites (voir Ghosh & Sen (1985) pour plus détails). Les travaux pionniers de Wolfe (1971) justifient l'usage de la simulation pour calculer la *p-valeur* de ce test.

4.1.5 Focus sur les mélanges de Logit dans le contexte des rachats

Les mélanges de régressions logistiques font partie intégrante des modèles mélanges semi-paramétriques. Un résumé et une revue bibliographique de ce type de modèles est disponible dans le papier de Lindsay & Lesperance (1995), et des applications dans le domaine de la biologie sont fournies dans Follmann & Lambert (1989) et Wang (1994). Nous avons vu au chapitre 2 que la régression logistique était adaptée aux données binomialement distribuées. Ainsi en notant p_{ij} la probabilité de rachat de N_j individus homogènes (ayant les mêmes caractéristiques) appartenant à la composante i du mélange, et en reprenant les notations ci-dessus avec $Y_j \sim Bin(N_j, p_i(X_j))$:

$$f(y_j; p_i(X_j)) = P(Y_j = y_j) = C_{N_j}^{y_j} p_i(X_j)^{y_j}(1 - p_i(X_j))^{N_j - y_j}, \tag{4.7}$$

où y_j est le nombre de rachats observés dans le groupe homogène j, et $p_i(X_j)$ résulte du lien logistique

$$p_i(X_j) = \frac{\exp(\beta_i^T X_j)}{1 + \exp(\beta_i^T X_j)},$$

avec $X_j = (X_{j1}, ..., X_{jk})^T$ le vecteur des k covariables de l'individu j, $\beta_i = (\beta_1, ..., \beta_k)^T$ le vecteur des k coefficients de régressions de la composante i.

Considérons la moyenne et la variance du modèle mélange de régressions logistiques donné par

$$f(y_j;\Psi)=\sum_{i=1}^{g}\pi_i(X_j^{'})f(y_j;p_i(X_j)).$$

L'interprétation de ce modèle est la suivante : il existe plusieurs groupes qui suivent des distributions logistiques différentes, avec chacun une proportion $\pi_i(X_j^{'})$. La moyenne et la variance sont facilement calculables par les formules suivantes :

$$\begin{aligned}
\mathbb{E}[Y_j] &= \mathbb{E}_Z\left[\mathbb{E}[Y_j \mid Z_j]\right]=\sum_{i=1}^{g}P(Z_{ij}=1)\mathbb{E}[Y_j \mid Z_j]=\sum_{i=1}^{g}\pi_{ij}p_{ij},\\
Var[Y_j] &= \mathbb{E}\left[Var[Y_j|Z_j]\right]+Var\left[\mathbb{E}[Y_j|Z_j]\right]\\
&= N_j\left[\sum_{i=1}^{g}\pi_{ij}p_{ij}\right]\left[1-\sum_{i=1}^{g}\pi_{ij}p_{ij}\right]+\frac{(N_j-1)}{N_j}Var\left[\mathbb{E}[Y_j|Z_j]\right],
\end{aligned}$$

avec $Var\left[\mathbb{E}[Y_j|Z_j]\right]=N_j^2\left[\sum_{i=1}^{g}\pi_{ij}p_{ij}^2-(\sum_{i=1}^{g}\pi_{ij}p_{ij})^2\right]$.

Dans notre cas, nous considérons également des poids $\pi_i(X_j^{'})$ qui dépendent de certains facteurs endogènes ou exogènes, ce qui nous amène à les définir aussi comme des régressions logistiques multinomiales (rappelons que Z_j est multinomiale), soit :

$$\pi_i(X_j^{'})=\frac{\exp(\gamma_i^T X_j^{'})}{\sum_{h=1}^{g}\exp(\gamma_h^T X_j^{'})}, \tag{4.8}$$

avec $X_j^{'}=(X_{j1}^{'},...,X_{jl}^{'})^T$ un ensemble de l covariables de l'individu j et $\gamma_i=(\gamma_1,...,\gamma_l)^T$ le vecteur des l coefficients de régression du poids de la composante i. Ainsi le vecteur des paramètres à estimer vaut $\Psi=(\gamma_1^T,...,\gamma_g^T,\beta_1^T,...,\beta_g^T)^T$. Pour effectuer l'estimation de Ψ, il suffit donc d'insérer (4.7) et (4.8) dans les formules de l'algorithme EM qui sont valables en toute généralité. Dans le cadre de mélange de régressions logistiques, le lecteur intéressé par les problèmes d'identifiabilité pourra trouver son bonheur dans les travaux de Margolin et al. (1989) et Teicher (1963), qui donnent des conditions nécessaires et suffisantes pour leur résolution. Nos futurs choix de modélisation satisfont ces conditions.

En ce qui concerne les prévisions de taux de rachat, elles sont calculées par aggrégation des décisions individuelles sur chaque pas de temps (les études seront trimestrielles). Ces décisions étant indépendantes, le principe de calcul de l'intervalle de confiance est identique à celui développé dans le chapitre 2. Pour connaître la décision individuelle d'un assuré, nous regardons les probabilités à-posteriori d'appartenir à chacune des composantes : selon la règle de Bayes, l'individu appartient à la composante pour laquelle la probabilité $\pi_i(X_j^{'})$ d'appartenance est maximale. Etant donné cette appartenance, nous calculons ensuite sa probabilité de rachat $p_i(X_j)$ associée à cette composante. Si une variable explicative est incluse dans le calcul des poids des composantes, alors la proportion de cette composante peut évoluer en fonction de l'évolution de cette variable (introduction d'une corrélation). Dans le cas classique où les poids ne dépendent d'aucune variable endogène ou exogène, un individu appartient à un seul et unique groupe au cours de la vie de son contrat.

4.2 Cas pratique d'utilisation de mélange de Logit

Nous utilisons dans cet exemple les mêmes données et la même méthodologie (leur description ayant déjà été faite) que dans l'analyse dynamique afin de construire le modèle mélange. La période d'apprentissage représente toujours les deux tiers de la période totale, sachant que nous validons le calibrage du modèle sur la période de validation. Nous exposons dans cette partie les différents résultats que nous retourne l'étude par mélange dans ce contexte ; à savoir une estimation des paramètres de chaque densité composante, une estimation des paramètres de chaque poids et leur robustesse, une comparaison du taux de rachat observé avec la projection par le modèle de ce taux sur l'échantillon de validation, et enfin un test de type Kolmogorov qui permet de valider ou non la qualité des prévisions. Nous commenterons ces résultats en y apportant une tentative de justification pratique. L'exposition de l'ensemble de ces résultats sera reprise dans le chapitre des applications, ce qui permettra d'énoncer la véritable découverte faite dans le cadre de ce travail sur la manière de prendre en compte les différents facteurs de risque pour diverses grandes familles de produits d'épargne en Assurance-Vie.

Considérons donc les produits mixtes du portefeuille espagnol pour lesquels nous avons déjà tenté une modélisation sans succès, même par l'introduction de variables financières et économiques (cf figure 3.1). La popularité des produits mixtes en Espagne n'est plus à démontrer. Comme déjà évoqué dans les chapitres précédents, le contrat mixte est un contrat d'épargne temporaire classique qui a l'avantage de retourner à son bénéficiaire un capital choisi lors de la souscription, et ce quelque soit l'état de l'assuré (en vie ou décédé, d'où l'appelation "mixte" de cette garantie). Ce type de contrat est très répandu sur le marché espagnol où il a rencontré un vif succès bien que son prix soit plus élevé qu'un pur contrat d'épargne, puisque le risque encouru par l'assuré est plus faible. Les variables disponibles et la période d'étude (1/1/2000 au 31/12/2007) restent inchangées. La construction des données qui vont nous servir à construire le modèle est identique hormis le pas de temps qui passe de mensuel à trimestriel, pour éviter une trop grande volumétrie de données (qui ne plaît pas beaucoup au logiciel R).

Rappelons les informations dont nous disposons dans la base d'origine : le numéro du produit, la date d'émission, la date de sortie et sa raison (si sortie il y a), l'option de participation aux bénéfices de l'entreprise (PB), la date de naissance de l'assuré, son sexe, sa richesse, la fréquence de la prime, la prime de risque et la prime d'épargne. Les variables que nous pouvons potentiellement insérer dans la modélisation mélange de régressions logistiques sont donc l'ancienneté du contrat (par la variable "duration.range"), la clause de participation aux bénéfices de la compagnie (renommée "PB.guarantee" mais anciennement "contract.type"), la tranche d'âge de souscription (par "underwritingAge.range"), la tranche de richesse de l'assuré (par "fa.range"), la fréquence de la prime (par "premium.frequency"), les valeurs de prime (de risque par "riskPrem.range" et d'épargne par "savingPrem.range"), et les variables d'environnement que sont l'IBEX 35 (indice boursier espagnol) et le taux des obligations d'Etat 10 ans (par "rate10Y"). Nous considérons plus exactement un historique arbitraire de ces variables économiques puisque nous regardons leur valeur à la date de rachat comparée à leur valeur trois mois auparavant (une option de nos programmes permet de modifier ce critère : allonger la période *delta* de regard en arrière, ou considérer la moyenne de cette évolution).

Nous présentons dans la suite quelques statistiques descriptives préalables sur la base de données des contrats mixtes, qui vont nous être fort utiles pour nous guider dans les choix de modélisation.

4.2.1 Analyse descriptive

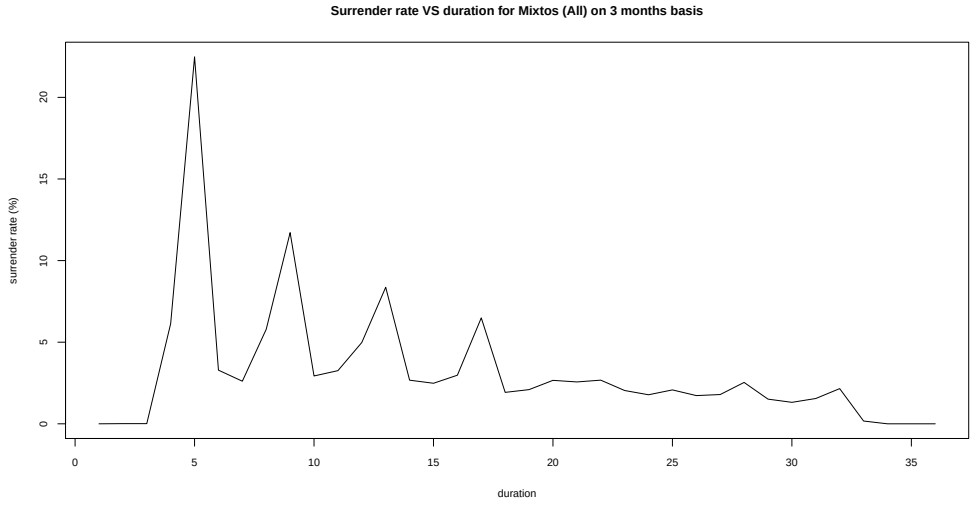
Evolution de l'exposition et du taux de rachat du portefeuille Nous constatons à travers l'étude du graphique 4.1 que la trajectoire du taux de rachat présente globalement trois phases : un plateau de niveau moyen de rachat bas entre 2000 et 2003, une hausse et une stabilisation entre 2004 et 2006, puis un pic de rachat suivi d'une chute vertigineuse du taux dans l'année 2007 malgré une exposition encore conséquente (bien que les nouvelles souscriptions soient rares à partir de 2006). Sur l'ensemble de la période, nous observons un phénomène périodique avec certains pics et creux de même amplitude qui semblent revenir régulièrement (même si estompé entre fin 2002 et fin 2003), d'où la pertinence de considérer une saisonnalité dans la modélisation (par la variable "month" ou "end.month" parfois).

FIGURE 4.1 – Exposition et taux de rachat trimestriel du portefeuille de produits Mixtes.

Profil des rachats par ancienneté de contrat La forme très spécifique du graphique 4.2 s'explique par les frais engendrés par des rachats entre les dates anniversaires du contrat, ce qui a déjà été discuté au chapitre 2. La catégorisation de l'ancienneté du contrat semble "obligatoire" pour rendre compte de l'aspect non-monotone de cette courbe, à moins de considérer un pas de temps annuel (et encore) qui ferait ressembler la courbe à une exponentielle décroissante. Ce choix de pas de temps annuel a été banni pour conserver un volume d'observations suffisant, nécessaire à la bonne construction et validation du modèle probabiliste.

Taux de rachat par cohorte Le taux de rachat global par cohorte pour les produits mixtes est tout à fait particulier et instructif car il représente le pourcentage d'assurés de la cohorte qui ont racheté leur contrat. Le graphique 4.3 montre clairement des comportements très hétérogènes entre cohortes. Au vu du graphique 4.2 qui présente une décroissance homogène des pics de rachat en fonction de l'ancienneté, il est difficile de comprendre pourquoi les anciennes cohortes (2000 à 2002) ont un taux global de rachat de l'ordre de 30 % alors que celles entre 2002 et 2006 s'approchent des 80 %. Le bas niveau de rachat des plus jeunes cohortes s'explique facilement par le fait que le rachat est interdit en première année de contrat. Cette hétérogénéité entre cohortes

FIGURE 4.2 – Rachat par ancienneté de contrat (en trimestre) pour les produits Mixtes.



a sûrement une explication rationnelle, toutefois difficile à récupérer même s'il s'agit probablement d'une politique de vente (ou législation) changeante sur ces produits.

Taux de rachat par date et par ancienneté de contrat La vision 3D du graphique 4.4 fournit une information additionnelle : ce n'est que récemment que le profil spécifique des rachats en fonction de l'ancienneté est apparu. Cette information primordiale vient valider la mise en place récente d'une spécificité (rachat sans frais aux dates anniversaires) qui a aussi pour effet l'apparition du créneau de la figure 4.3. L'hétérogénéité vient du mélange en portefeuille de ces deux types de population, car les changements n'ont visiblement été effectifs que pour les nouvelles souscriptions à partir de 2002.

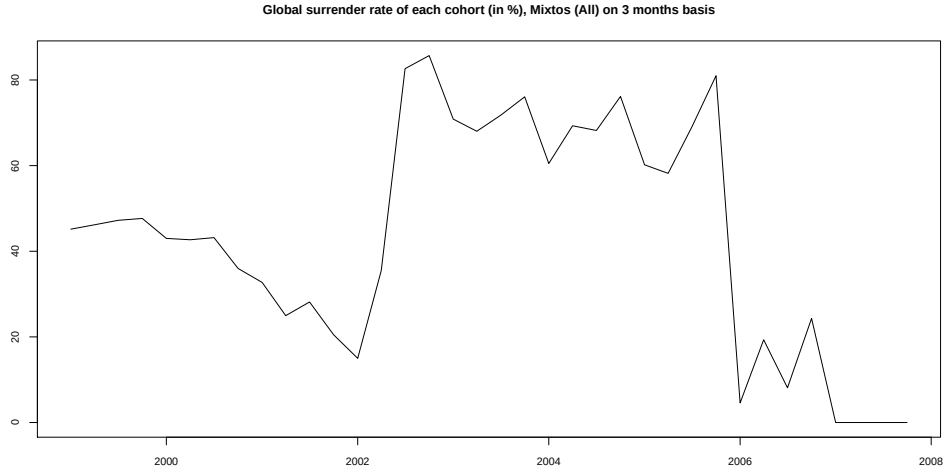
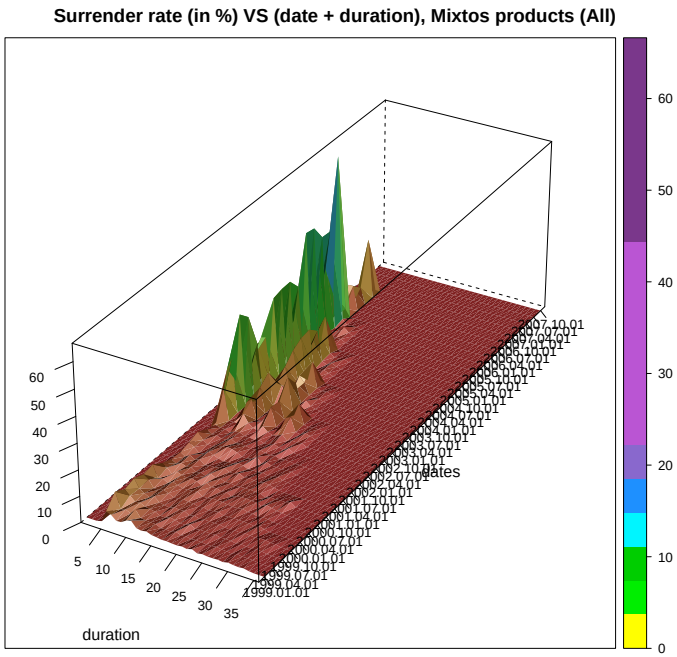


FIGURE 4.3 – Pourcentage global de rachat par cohorte pour les produits Mixtes.

FIGURE 4.4 – Profil 3D du taux de rachat par date et par ancienneté de contrat (par trimestre), produit Mixtes.



4.2.2 Sélection des variables par CART

Taux d'erreur de classification de l'arbre Le classifieur par forêts aléatoires avec en variables d'entrée l'ensemble des variables à disposition (et pas seulement les variables catégorielles) donne d'excellents résultats. Le taux d'erreur de la matrice de confusion est de 4,6 %, avec une sensibilité de 99,5 % et une spécificité de 84 %. Ces statistiques nous sécurise quant au classement (énoncé dans le paragraphe suivant) du pouvoir discriminant des variables.

	Rachats non-observés	Rachat observés
Rachats non-prédits	4599	877
Rachats prédits	85	15485

Importance des variables explicatives Comme énoncé dans la section 2.3.1, nous avons vérifié que le classement de l'importance des variables explicatives soit le même pour les périodes de pics de rachat comme pour les périodes creux (ce qui est le cas) lorsque nous regardons les comportements de rachat en fonction de l'ancienneté du contrat. Il n'y a donc pas de biais introduit dans les résultats de la figure 4.5, qui va nous servir de base pour la prise en compte des bons inputs lors de la modélisation. Nous prenons ainsi en priorité les variables de saisonnalité et d'ancienneté de contrat (catégorisée) déjà validées comme importantes, en y ajoutant l'option de PB et la prime de risque (catégorisée également car relation non-monotone encore une fois).

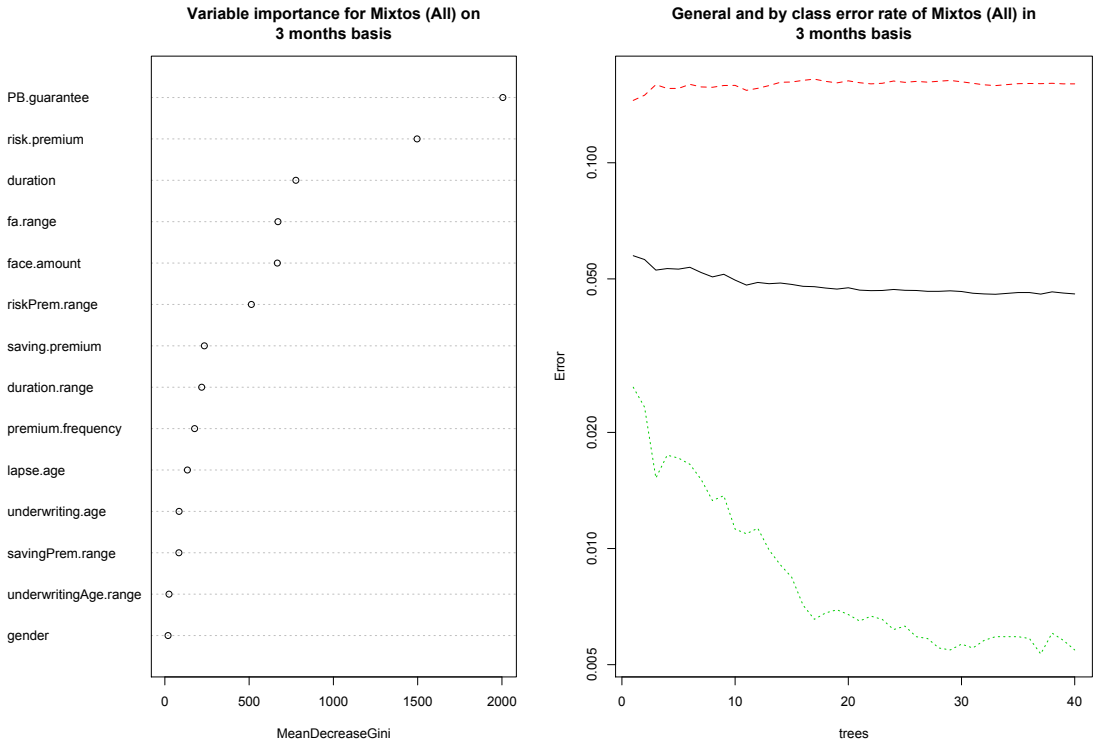


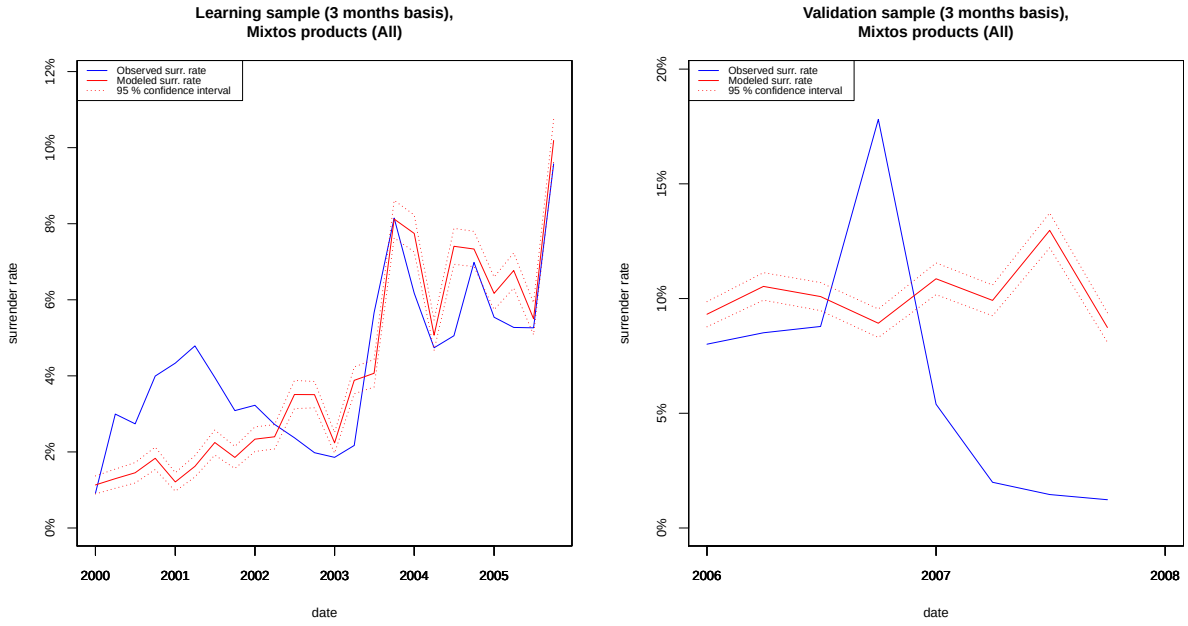
FIGURE 4.5 – Importance des variables explicatives, produit Mixtes.

4.2.3 Modélisation et prévisions par mélange de GLM

Nous ne commentons pas de nouveau le graphe 4.6 car cela a été fait longuement en section 3.1. Le pas trimestriel implique cependant une moins bonne estimation de la réalité sur la période d'apprentissage (en comparaison avec la figure 3.1), tout en conservant le problème majeur du changement de niveau du taux de rachat en 2007 qui n'est pas prévu par le modèle. Parmi des modèles mélange de régressions logistiques de deux à cinq composantes, nous choisissons celui qui minimise le critère BIC de sélection de modèle. Rappelons juste que ce critère prend en compte la complexité du modèle en pénalisant la vraisemblance d'un modèle qui contiendrait beaucoup de paramètres à estimer. Pour les produits mixtes, le modèle retenu a cinq composantes (nous verrons dans le chapitre suivant que ce n'est heureusement pas toujours le modèle avec le plus de composantes qui est retenu !), traduisant ainsi une forte hétérogénéité des données. Les résultats probants de la courbe 4.7 renforcent l'adéquation de la modélisation par mélange pour ce type de produit. En effet, nous constatons inévitablement le "super" pouvoir prédictif de la méthode qui s'ajuste quasi-parfaitement avec malgré un intervalle de confiance étroit, garantissant la robustesse de la modélisation.

Impact des variables explicatives par les mélanges de Logit Nous avons vu comment choisir (CART) les variables explicatives à entrer dans la modélisation, de même que l'apport théorique des modèles mélange qui vont nous permettre de tenir compte de la forte hétérogénéité des comportements mise en évidence grâce à des statistiques descriptives ciblées. Il est maintenant grand temps de dévoiler notre intuition, celle qui nous guidera dans la modélisation de l'ensemble des familles de produit jusqu'à

FIGURE 4.6 – Modélisation et prévision du taux de rachat des produits Mixtes par régression logistique dynamique.



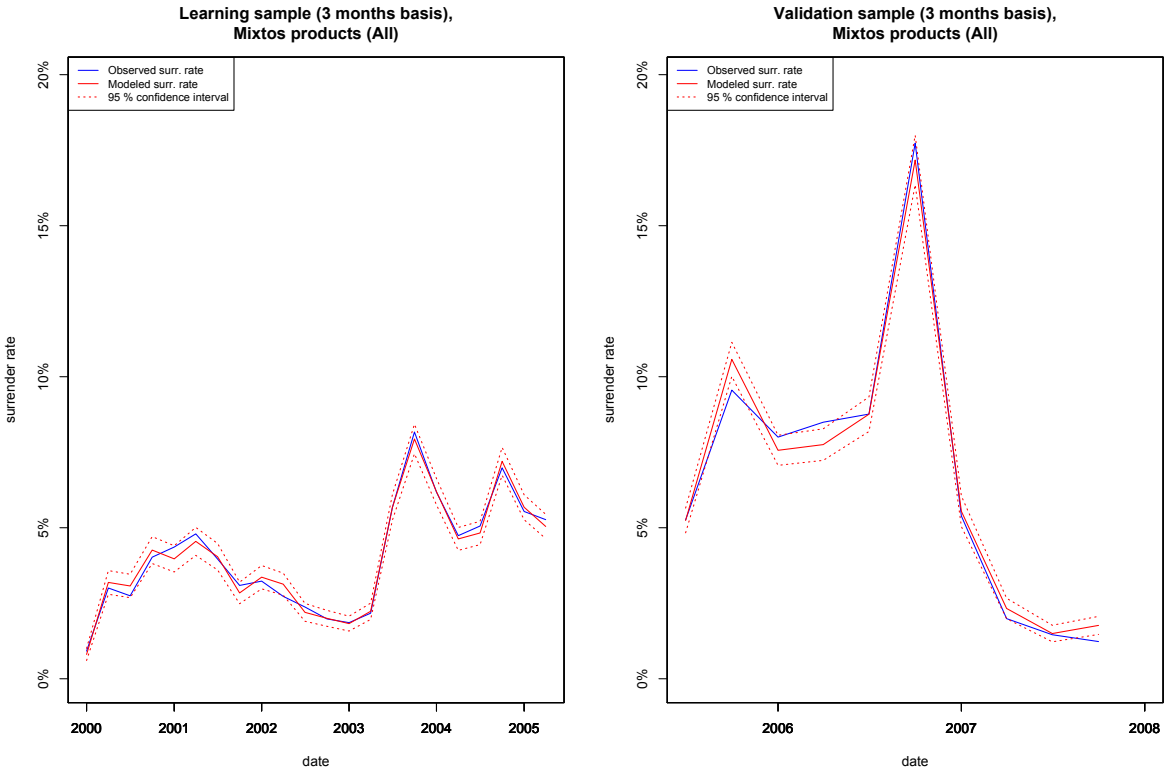


FIGURE 4.7 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Mixtes.

la fin de ce mémoire. Selon nous, la logique voudrait que les effets structurels ne soit pas une source d'hétérogénéité entre comportements car ils s'appliquent à l'ensemble de la population sans distinction apparente. En exemple, une contrainte fiscale reste une contrainte valable pour tous les assurés du portefeuille, et ce quelque soient leurs autres caractéristiques. Par conséquent l'idée est de fixer les coefficients de régression constant entre les composantes pour les facteurs de risque associés à ces effets structurels, ce qui permet notamment de fortement limiter le nombre de paramètres à estimer. En revanche, la mixité de la population en termes de richesse et de sensibilité par rapport à l'environnement externe invoque des comportements de rachat totalement différents car très personnels. Il va ainsi se créer des groupes d'assurés pour lesquels il n'y a aucune raison que les paramètres de régression associés aux effets conjoncturels aient la même valeur, de même que pour le risque de base représenté par l'ordonnée à l'origine ("intercept"). Cette proposition fort logique et simple nous permet de prendre en compte une bonne part des différents comportements et de reconstruire correctement l'historique (en *back-testing*) du taux de rachat. L'estimation des coefficients de régression du modèle mélange appliqué aux contrats mixtes est disponible en figure 4.8. Les impacts des facteurs de risque sont les suivants :

– effets *structurels* (identiques quelque soit la composante d'appartenance) : un faible effet de saisonnalité est détecté (valeur absolue des coefficients de régression associés faible) avec globalement de plus en plus de rachats en approchant de la fin

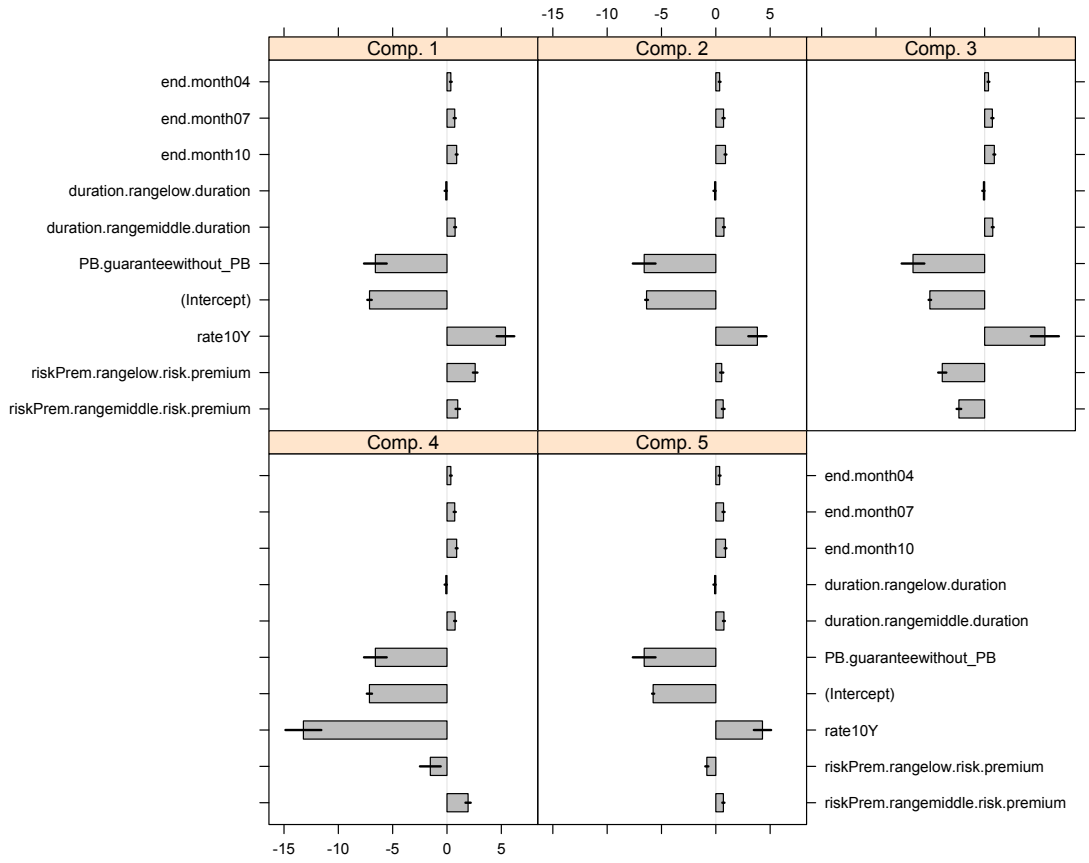


FIGURE 4.8 – Coefficients de régression des composantes du mélange de Logit.

de l'année civile. L'effet de l'ancienneté du contrat catégorisée en trois modalités ("low", "middle" et "high") est clair : c'est dans la tranche des anciennetés moyennes que se situent le plus grand nombre de rachats, et globalement les assurés rachètent assez rapidement. Le fait de ne pas avoir l'option de PB dans son contrat vient fortement diminuer la probabilité individuelle de rachat. Ces résultats viennent confirmer les observations faites auparavant.

– effets *conjoncturels* : l'effet de la prime de risque (liée à la richesse des assurés) est hétérogène et semble dicter en partie la sensibilité des agents aux mouvements de l'économie. Cela rejoint l'idée exprimée par les équipes Marketing qui disaient que la réactivité des assurés est relative à ce qu'ils possédent. Nous pouvons remarquer que les groupes pour lesquels le niveau de richesse est fortement discriminant réagissent en plus forte proportion aux mouvements du taux 10 ans.
– effets de *corrélation* : introduite via le contexte économique. Le taux 10 ans a une importance prépondérante en termes d'impact (valeur absolue du coefficient associé élevée). La calibration montre que certains assurés rachètent plus lorsque le taux long-terme augmente (composantes 1, 2, 3 et 5) alors que les autres adoptent le comportement inverse, la rationalité étant illustrée par une augmentation des rachats en cas de hausse des taux sur ce type de produit (rendement garanti). Chaque trimestre, la hausse ou la baisse de ce taux va augmenter ou baisser **en même temps** la probabilité de rachat des assurés appartenant à un groupe donné, faisant évoluer dynamiquement la distribution du mélange au cours du temps.

Ces calibrations semblent robustes (l'écart-type des estimations est représenté par une proportion de la petite barre noire) puisque la valeur nulle n'appartient à aucun

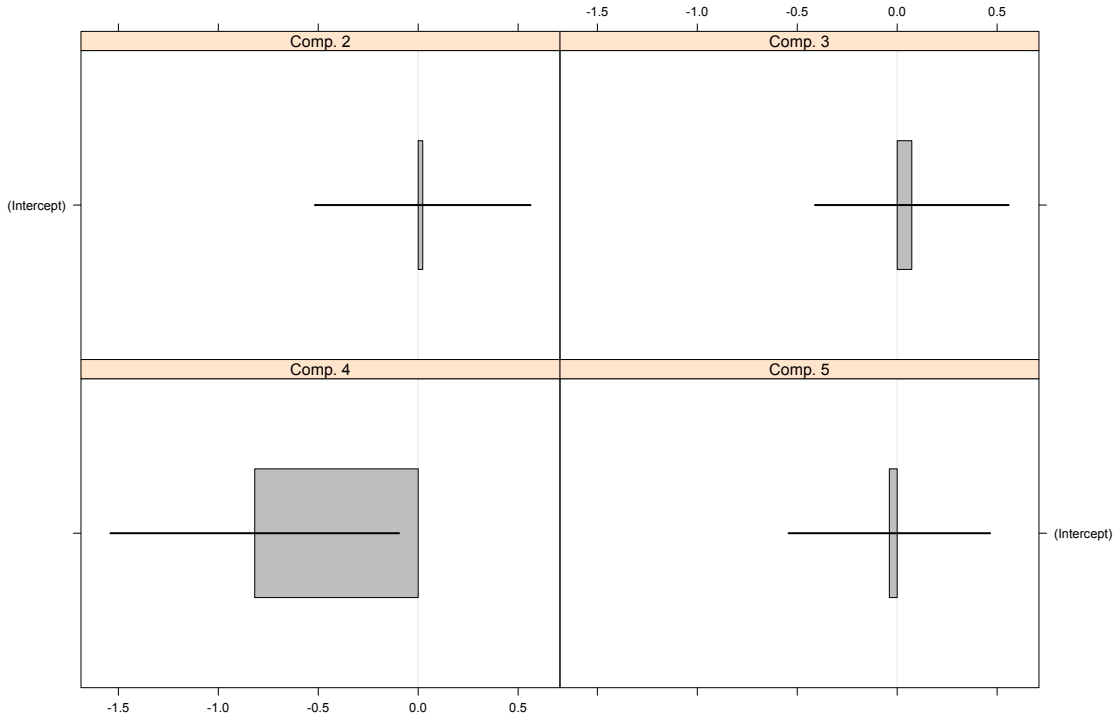


FIGURE 4.9 – Coefficients de régression des poids des composantes du mélange de Logit.

intervalle de confiance de ces estimations. Pour ce qui est de la calibration des poids de chaque composante, les résultats semblent moins robustes et ce sera souvent le cas en pratique. La figure 4.9 résume les proportions de chaque composante dans le mélange (pas de variable explicative ici), nous obtenons : par la formule (4.8)

$$\pi_1 = 22\%, \pi_2 = 23\%, \pi_3 = 24\%, \pi_4 = 10\%, \pi_5 = 21\%,$$

ce qui prouve qu'il n'y a pas de composantes inutiles, chacune ayant son importance dans le mélange.

Pour vérifier la robustesse de cette approche autrement que par l'aspect visuel, nous appliquons deux tests : un test de normalité des résidus (Pearson), et un test sur les distributions (Wilcoxon Mann-Whitney). Nous ne détaillons pas le test de Pearson qui est un des plus connus ; le principe du test de Wilcoxon-Mann-Whitney est donné ci-dessous. Les résultats de ces deux tests pour un seuil de 5 % suivent dans le tableau 4.1. Nous ne pouvons donc pas rejeter l'hypothèse nulle qui correspond au fait que la variable aléatoire "observée" et la variable aléatoire "prédite" aient la même distribution. Les sorties R des résultats numériques de ces tests sont disponibles en annexe D.1.

	Test de Pearson	Test de Wilcoxon-Mann-Whitney
p-valeur	0.8495	0.7394

TABLE 4.1 – p-valeur des tests de résidus et de distribution pour validation.

Idée du test de Wilcoxon-Mann-Whitney Supposons deux lois P_x et P_y inconnues. Nous rassemblons les deux échantillons que sont les valeurs observées et les valeurs prédites sur la période de validation, et nous les ordonnons. Si l'alternance des X_i et des Y_j est assez régulière, alors nous pouvons penser que les deux lois ont sensiblement la même distribution (qui est l'hypothèse nulle). Dans le cas contraire, nous pouvons douter de cette hypothèse. C'est donc un test basé sur les rangs, qui a ainsi l'avantage d'être non-paramétrique.

4.3 Conclusion

Nous proposons dans ce chapitre une méthodologie de prise en compte des facteurs de risque. La clef réside dans la distinction entre les effets structurels supposés constants entre groupes homogènes (d'un point de vue comportemental) d'assurés, et les effets conjoncturels qui sont autorisés à varier entre groupes. Cette suggestion provient d'une intuition logique et donne des résultats plus qu'acceptables dans un contexte de contrats mixtes. L'introduction des modèles mélange nous a permis non seulement d'élargir notre champs de connaissance mais aussi d'améliorer la flexibilité de la modélisation en permettant la représentation d'une éventuelle multimodalité de la densité des comportements de rachat, caractéristique d'une forte hétérogénéité. Notre prochain objectif est de tester la validité de cette méthodologie sur un large panel de type de contrat.

Chapitre 5

Application au portefeuille espagnol d'AXA

L'intérêt de cette partie réside dans l'application pratique des théories développées dans les chapitres précédents, dans le but de valider la méthodologie adoptée. Le portefeuille d'Assurance-Vie épargne d'AXA Seguros est utilisé dans toute sa "largeur", avec des résultats allant de produits de pure investissement à des produits alliant des composantes épargne à des garanties de prévoyance, en passant par des produits directement indexés sur les marchés financiers. Nous verrons que la modélisation proposée a un fort pouvoir d'adaptation et fournit des résultats très encourageants en termes de pouvoir prédictif, tout en conservant l'originalité de ne pas impliquer trop de facteurs explicatifs afin de ne pas trop complexifier le modèle. Chaque section de ce chapitre correspond à l'étude d'une famille de produit, avec toujours le même plan d'étude : une explication très succinte du type de contrat (car les produits sont aggrégés), une analyse simplifiée basée sur quelques statistiques descriptives, les résultats de la méthode CART, et les deux modélisations logistiques avec prévisions et tests associés. D'un point de vue granularité des données, il est nécessaire d'étudier les rachats par famille de produits au maximum (une aggrégation encore plus grande n'aurait plus de sens) car les supports d'investissement et les options classiques varient d'une famille à l'autre, ce qui apporte des changements importants en termes de modélisation. Il va sans dire que l'idéal est d'affiner les études à l'échelle de lignes de produits, voire de produits. L'outil informatique que nous avons développé permet de choisir son niveau de granularité, mais nous préférions montrer que notre méthode fonctionne à une échelle d'aggrégation importante (sachant qu'à l'échelle d'un produit, cette modélisation est souvent moins complexe car nous connaissons exactement toutes les clauses et options qui impactent le rachat et il suffit de les inclure dans la modélisation). De plus, une étude par produit ne permettrait pas de modéliser les rachats globalement, car les corrélations entre produits sont très difficilement calibrables.

Pour les résultats de la modélisation par mélange de régressions logistiques, nous avons choisi de commenter les effets des variables explicatives au vu des estimations des coefficients de régression sans pour autant afficher les "boxplots" correspondants pour des soucis de concision. Le lecteur intéressé pourra consulter les annexes E pour accéder à ces informations plus précises. Toute l'étude est basée sur un pas de temps trimestriel et sur une période de retour *delta* (longueur de temps depuis la date de rachat sur laquelle l'assuré regarde la performance des indices économiques) de un trimestre, ces options pouvant être ajustée dans notre outil (pas mensuel, trimestriel ou annuel et *delta* doit être un entier positif).

5.1 Les contrats de pure investissement (Ahorro)

Les contrats "Ahorro" sont des contrats de pure épargne. Ils offrent un rendement différent suivant le produit considéré, mais tous sont des taux garantis (le risque de taux est donc porté par l'assureur). Nous pourrions comparer ces contrats à des contrats bancaires, avec la différence qu'ils offrent des avantages fiscaux et/ou des garanties supplémentaires. Les informations dont nous disposons pour ce type de contrats sont le numéro du produit, la date d'émission, la date de sortie et sa raison (si sortie il y a), l'option de participation aux bénéfices de l'entreprise (PB), la date de naissance de l'assuré, son sexe, sa richesse, la fréquence de la prime, la prime de risque et la prime d'épargne. Un aperçu des données formatées est disponible en annexe E.1.1. La période de données va de début 1999 à fin 2007 (certains contrats sont évidemment souscrits avant 1999), mais la période d'étude s'étend du 1/1/2000 au 31/12/2007 car les rachats n'ont été répertoriés qu'à partir de début 2000.

5.1.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Nous représentons dans la figure 5.1 l'historique de l'exposition (en vert), du taux de rachat (en noir) et du taux de chute (en rouge) du portefeuille. Nous voyons bien que les rachats font partie des chutes, mais que les chutes englobent d'autres événements ; ici par exemple un produit largement distribué est arrivé à maturité début 2005. Nous observons une forte baisse du taux de rachat dans l'année 2007, le niveau moyen de rachat ayant diminué fin 2001 pour rester relativement stable ensuite (peu de volatilité jusqu'en 2007). Le taux de rachat semble présenter des creux et des pics réguliers traduisant une certaine périodicité, sans doute dûe au cycle annuel de vente des produits (période de fête,...).

Profil des rachats par ancienneté de contrat Le profil des rachats en fonction de l'ancienneté des contrats est un élément clef de modélisation. En effet, il s'agit ici de détecter un éventuel aspect monotone de la courbe afin de savoir si la catégorisation de la variable "ancienneté" serait judicieuse. En figure 5.2, des pics périodiques apparaissent

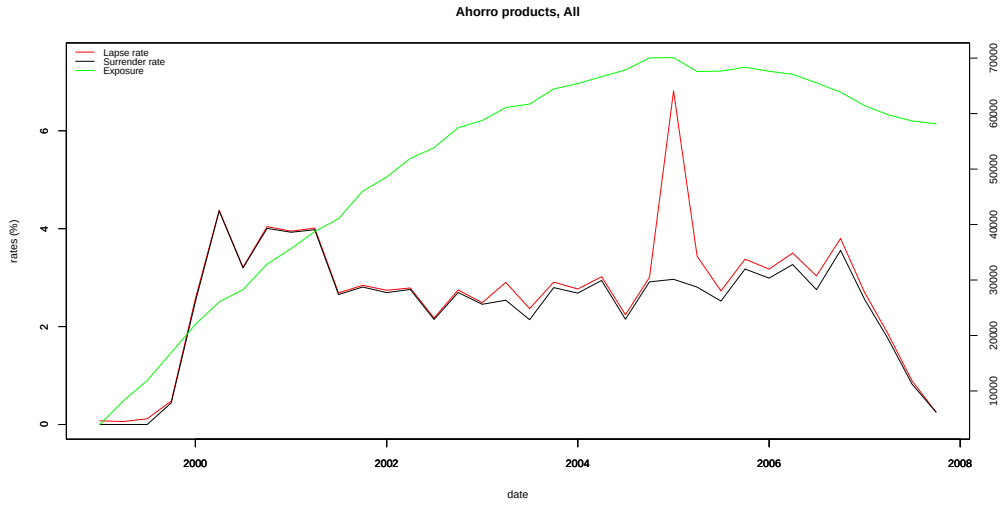
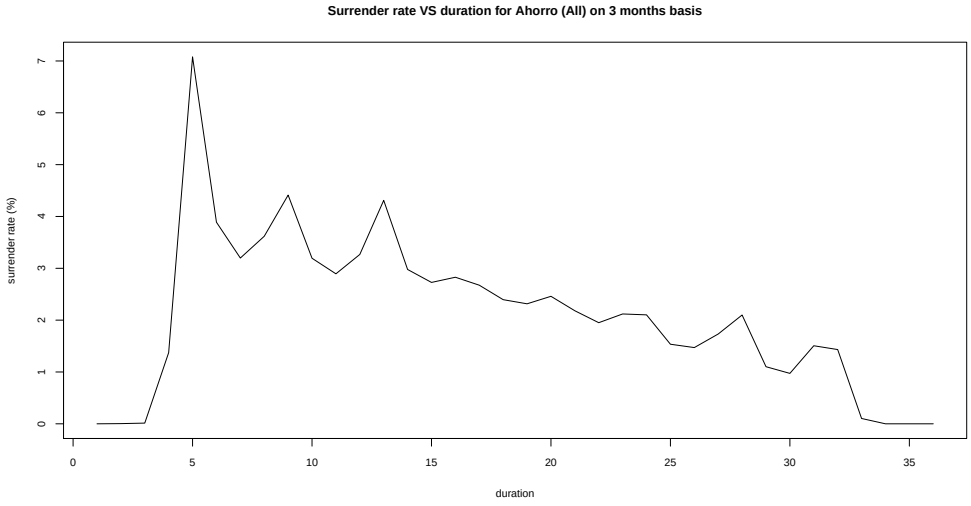


FIGURE 5.1 – Exposition et taux de rachat trimestriel du portefeuille de produits Ahorro.

FIGURE 5.2 – Rachat par ancienneté de contrat (en trimestre) pour les produits Ahorro.

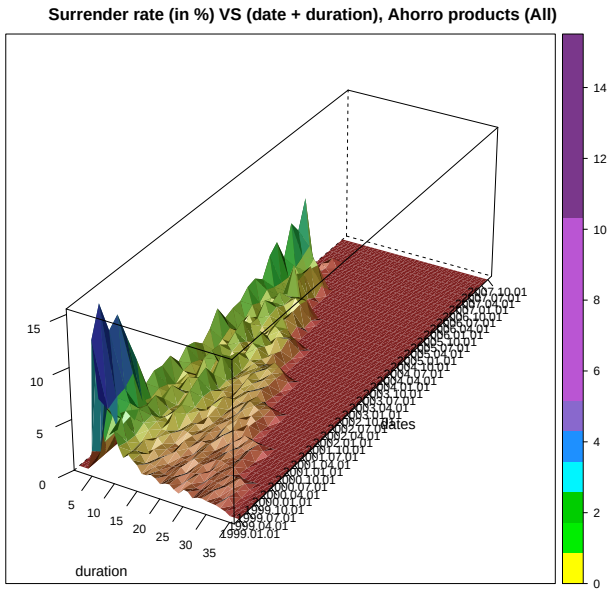


à chaque date anniversaire du contrat : cela est dû au fait que les contrats "Ahorro" en Espagne sont rachetables sans frais à chaque anniversaire de la police (aucun rachat n'est autorisé la première année, sauf cas exceptionnel). Ce profil suggère la catégorisation de cette variable continue, dans le sens où un unique coefficient de régression serait insuffisant à rendre compte de cette forme spécifique. Chaque fois qu'une catégorisation de variable continue sera effectuée dans la suite, ce sera par la méthode des quantiles : trois modalités avec chacune la même exposition (l'ancienneté était catégorisée différemment au départ, aussi il est possible qu'un résultat soit basé sur cette ancienne catégorisation mais ceci est marginal). Les pics de rachat s'amenuisent avec le temps pour la simple et bonne raison que l'exposition devient moindre.

Taux de rachat par cohorte Lorsque l'on regarde le taux de rachat global par cohorte, l'idée est de voir si certaines cohortes ont globalement beaucoup plus racheté que d'autres. La figure E.1 en annexe E.1.2 permet de détecter une partie de l'hétérogénéité des comportements susceptible d'exister : dans ce cas précis, rien ne semble anormal (c'est pourquoi nous basculons ce graphe en annexe), le taux chutant à 0 pour les très jeunes cohortes car personne n'a encore racheté (les assurés sont dans leur première année de contrat). Nous retrouvons d'ailleurs les caractéristiques de la figure 5.2 à travers les baisses périodiques observées.

Taux de rachat par date et par ancienneté de contrat La vision 3D offerte par la figure 5.3 est utile dans un contexte global. Il est relativement facile d'observer des comportement anormaux en croisant les effets des dates et de l'ancienneté du contrat. Ici par exemple, nous observons que les assurés rachètent majoritairement avant leur quatrième année de contrat (du trimestre 4 au trimestre 12) quelque soit la date ; bien qu'en 2000 énormément de personnes rachetaient dès le premier anniversaire de la police (pics bleus). C'est typiquement le mélange de ces comportements qui donne une surdispersion des données et qui empêche la modélisation par une approche simplifiée.

FIGURE 5.3 – Profil 3D du taux de rachat par date et par ancienneté de contrat (par trimestre), produit Ahorro.



5.1.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre Au vu de la matrice de confusion sur l'échantillon de validation, le taux d'erreur de classification de l'arbre s'élève à 6,77 %. Ce bon résultat est à prendre avec précaution car la spécificité est assez mauvaise (34 %), bien que la sensibilité soit excellente (99,6 %). Nous nous servons de ce classifieur relativement précis pour en extraire les variables discriminantes dans le paragraphe suivant.

	Rachats non-observés	Rachat observés
Rachats non-prédits	1731	3363
Rachats prédits	196	47281

Importance des variables explicatives Les variables qui apparaissent comme les plus discriminantes dans la figure 5.4 sont l'ancienneté de contrat, suivie de la richesse de l'assuré, de l'option de participation aux bénéfices (PB), de la prime d'épargne (corrélée à la richesse, donc nous ne considérerons qu'une des deux variables dans la modélisation), de l'ancienneté de contrat catégorisée (celle que nous considérerons par la suite pour mieux refléter le profil spécifique des rachats vu au graphe 5.2) et ainsi de suite. Ce classement nous sert de base dans le choix des inputs aux futures modélisations, sachant qu'il confirme quasiment tout le temps les statistiques descriptives du taux de rachat en fonction de ces variables explicatives (nous nous abstenons donc dans le mémoire d'exposer l'ensemble des statistiques descriptives des rachats en fonction de chaque variable, ce qui serait long et fastidieux). La relation entre le taux de rachat et les variables explicatives continues n'étant que très rarement monotone, nous considérons très souvent dans la suite le classement par importance des variables catégorisées. Les trois principales que nous retenons ici sont donc l'option de PB, l'ancienneté de contrat et la fréquence de la prime. La saisonnalité n'apparait pas car elle ne fait pas partie des

variables en input de la méthode CART mais nous la prendrons toujours en compte, hormis avec des produits pour lesquels cet effet semble peu logique (produits structurés par exemple).

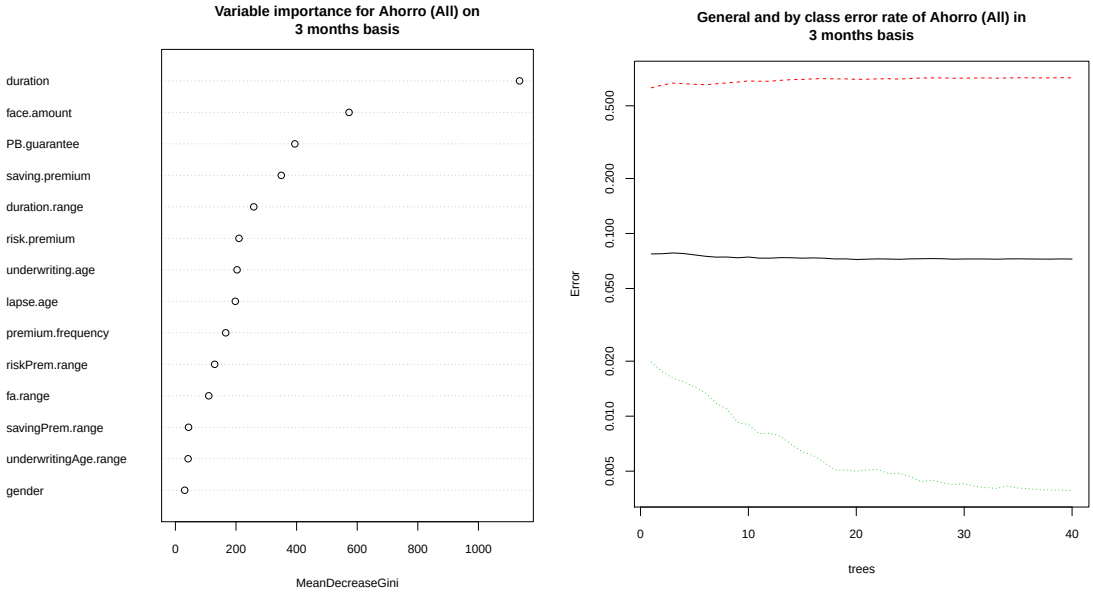


FIGURE 5.4 – Importance des variables explicatives, produit Ahorro.

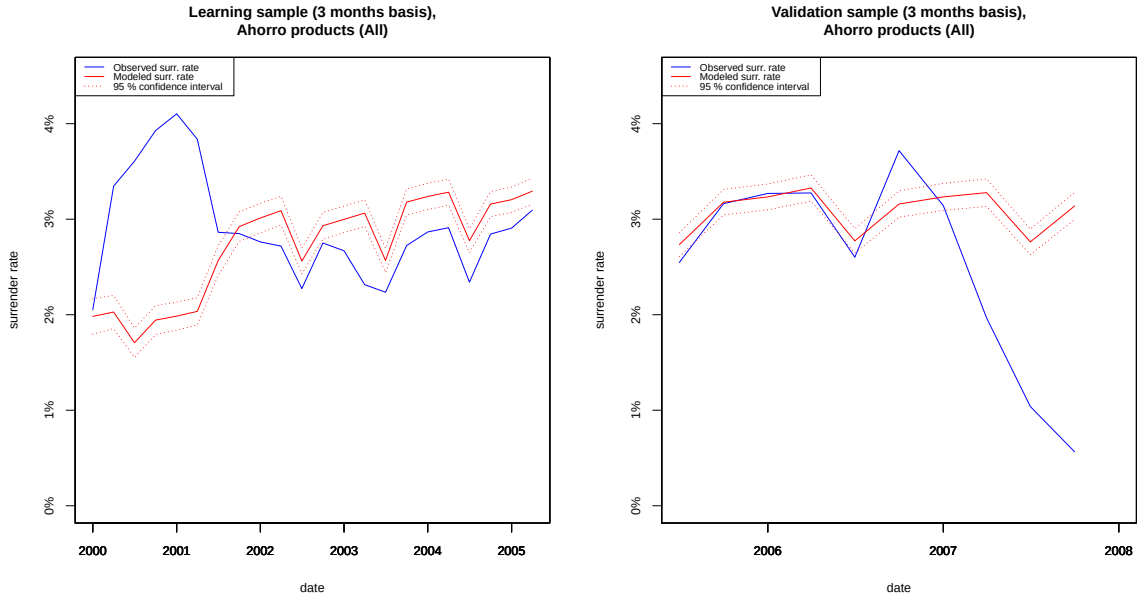
5.1.3 Modélisation et prévisions par mélange de GLM

Pour toutes les applications suivantes, les mêmes variables explicatives sont considérées en input de la modélisation dynamique et de la modélisation par mélange. L'approche par mélange permet de prendre ces variables en compte de manière différente, mais il est primordial de garder à l'esprit que nous prenons exactement les mêmes informations en entrée des modèles afin de comparer ce qui est comparable. Cette remarque justifiera le fait que certains modèles mélange ne sont pas optimisés (en termes de variables considérées, de nombre de composantes car parfois certaines composantes se ressemblent fortement...). Dans une optique où la volonté de l'utilisateur est de trouver la meilleure solution de modélisation, cette optimisation est tout à fait réalisable dans des délais raisonnables.

Le but est de comparer l'approche par mélange de régressions logistiques avec la régression logistique dynamique, et de voir s'il y a un apport conséquent de cette nouvelle modélisation. Nous discutons de l'impact des facteurs de risque suivant les groupes d'assurés dans le cadre de la modélisation mélange, et effectuons des comparaisons grâce aux prévisions des décisions individuelles qui nous permettent de reconstruire le taux de rachat par date.

Comparaison et discussion Les résultats de la régression logistique dynamique simple sont très frappants tellement ils sont mauvais (graphe 5.5). La cause de cette "faillite" est l'environnement économique changeant qui est mal modélisé, pour preuve la valeur du coefficient de régression consacré à l'impact du taux 10Y qui est extrêmement faible (0,06). Cela signifie qu'une forte variation de ce taux n'a que peu d'impact sur la probabilité finale de décision individuelle de rachat, ce qui est évidemment très

FIGURE 5.5 – Modélisation et prévision du taux de rachat des produits Ahorro par régression logistique dynamique.



discutable. Nous constatons également que le modèle logistique dynamique modélise bien la périodicité.

De par la flexibilité permise par les mélanges, les prévisions s'avèrent nettement plus justes et précises aussi bien sur la période d'apprentissage que sur la période de validation (graphe 5.6). Ce changement se retrouve notamment dans la valeur des coefficients de régression correspondant au taux 10Y (entre 10 et 100 plus élevé suivant

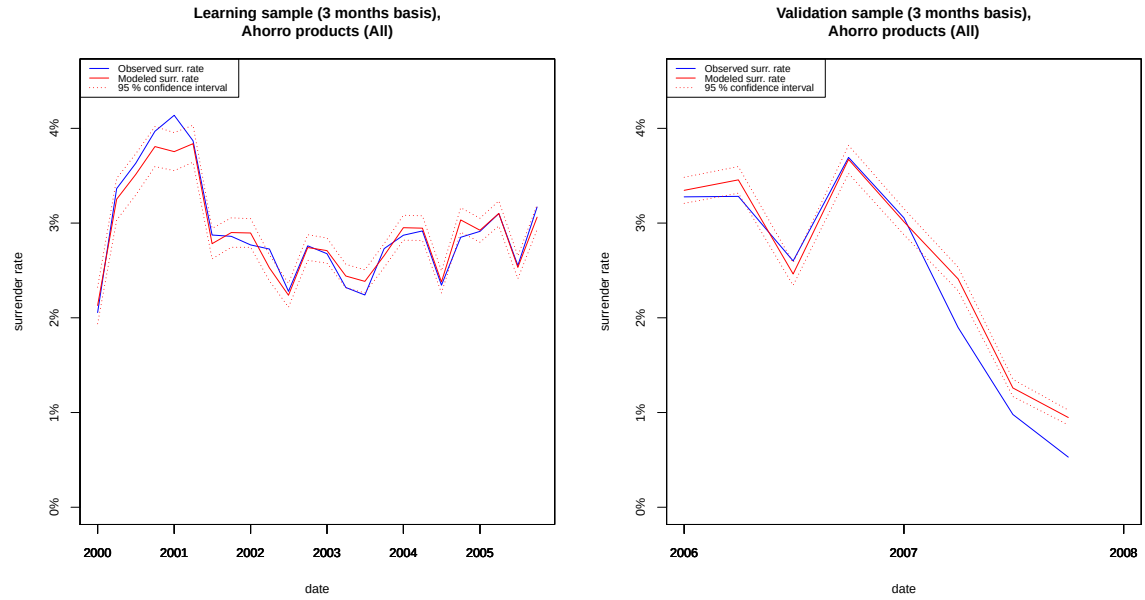


FIGURE 5.6 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Ahorro.

les composantes), traduisant un impact nettement plus réaliste de cette variable (voir figure E.2).

Impact des variables explicatives par les mélanges de Logit Nous partons du postulat que l'hétérogénéité provient de facteurs de risque qui peuvent avoir un effet différent suivant les personnes. L'idée de base est donc que les effets structurels bien connus (ancienneté de contrat, saisonnalité) sont censés avoir un impact homogène et constant quelque soit les groupes d'assurés considérés, alors que les effets conjoncturels (environnement économique) jouent différemment suivant les assurés. La mise en oeuvre de cette idée requiert de spécifier une estimation identique des coefficients de régression correspondant aux effets structurels pour toutes les composantes, en permettant aux coefficients de régression dédiés aux effets conjoncturels de varier entre composantes. Les professionnels ont coutume de considérer un taux d'intérêt long terme pour les produits de pure épargne à rendement garanti, aussi nous avons pris le taux 10 ans (taux 10Y). C'est ainsi que nous obtenons après estimation du modèle les coefficients de régression donnés en annexe E.1.2. Détaillons maintenant les impacts respectifs des facteurs de risque :

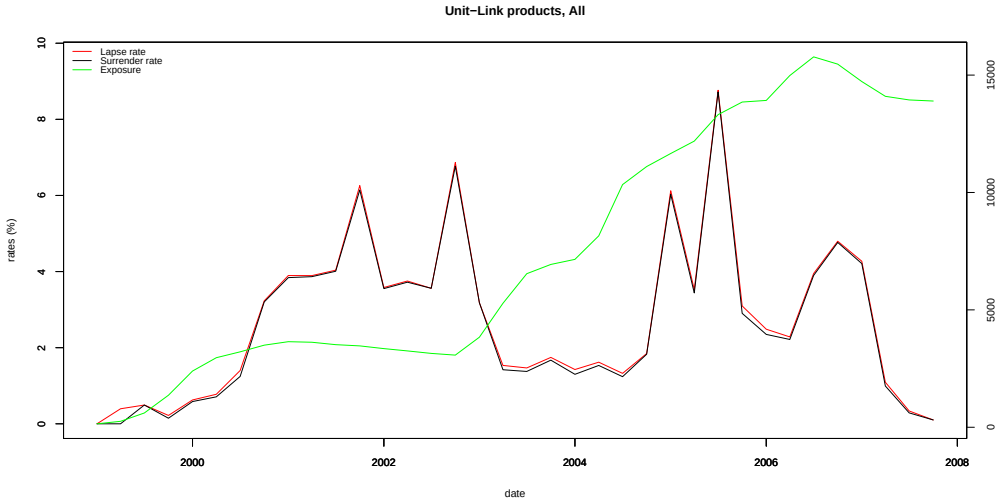
- effets structurels : identiques à tout le monde. En ce qui concerne la saisonnalité, moins de rachats constatés en été et environ le même taux de rachat en début et en fin d'année civile. Le risque de rachat est fort lorsque l'ancienneté de contrat (catégorisée en 3 modalités : faible, moyenne et longue) est faible, ce qui confirme les pics constatés dans le graphe 5.2. Plus les assurés sont âgés et moins la probabilité de rachat est grande, et l'effet de la fréquence de la prime (regroupée en 3 modalités : haute périodicité, moyenne et prime unique) est confirmé : plus la prime est fréquente et plus la probabilité de rachat est grande. Le fait de ne pas avoir l'option de PB abaisse fortement la probabilité de rachat.
- effets conjoncturels : deux groupes se distinguent. Pour le premier groupe, un taux 10Y qui augmente fait baisser la probabilité de rachat des assurés (composantes 1 et 3) alors que l'effet est inverse pour les autres groupes d'assurés. L'intensité de cette sensibilité caractérise ensuite les différentes composantes.

Nous constatons ainsi que les assurés réagissent différemment aux mêmes mouvements des effets du marché obligataire, venant confirmer l'irrationnalité et l'hétérogénéité des réactions.

5.2 Les contrats en Unités de Compte (Unit-Link)

Les contrats en UC sont des contrats qui offrent un rendement variable suivant les performances des marchés financiers. La rentabilité n'est donc pas garantie, bien qu'on adosse à certains de ces contrats des garanties plancher, ce qui limite le risque porté cette fois-ci par l'assuré. En général, ces contrats d'épargne offre des garanties supplémentaires telles qu'une couverture contre le décès, et les unités de compte sont basées sur des obligations et actions de diverses entreprises. Les informations dont nous disposons pour ce type de contrat sont le numéro du produit, la date d'émission, la date de sortie et sa raison (si sortie il y a), l'option de participation aux bénéfices, la date de naissance de l'assuré, son sexe, sa richesse, la fréquence de la prime, la prime de risque et la prime d'épargne. En fait le type d'information est identique que pour la famille précédente car nous avons la même base de données originelle, les donnés formatées ont donc le même aperçu (annexe E.1.1). La période d'étude est 1/1/2000 - 31/12/2007.

FIGURE 5.7 – Exposition et taux de rachat trimestriel du portefeuille de produits UC.



5.2.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille D'après le graphique 5.7, le niveau moyen du taux de rachat change régulièrement et fait preuve d'une volatilité assez importante. Les changements de niveau sont brusques et de forte amplitude, l'exposition se stabilise lors des périodes de crise (2000 et 2007), traduisant la réticence des agents à souscrire de nouvelles affaires sur ce type de produit en environnement fortement incertain. Nous retrouvons une forme de périodicité à travers les petites hausses et baisses régulières du taux, mais qui n'est pas non plus forcément évidente. Comme pour les produits de pure épargne, le taux de rachat semble s'effondrer à des niveaux anormalement bas en 2007.

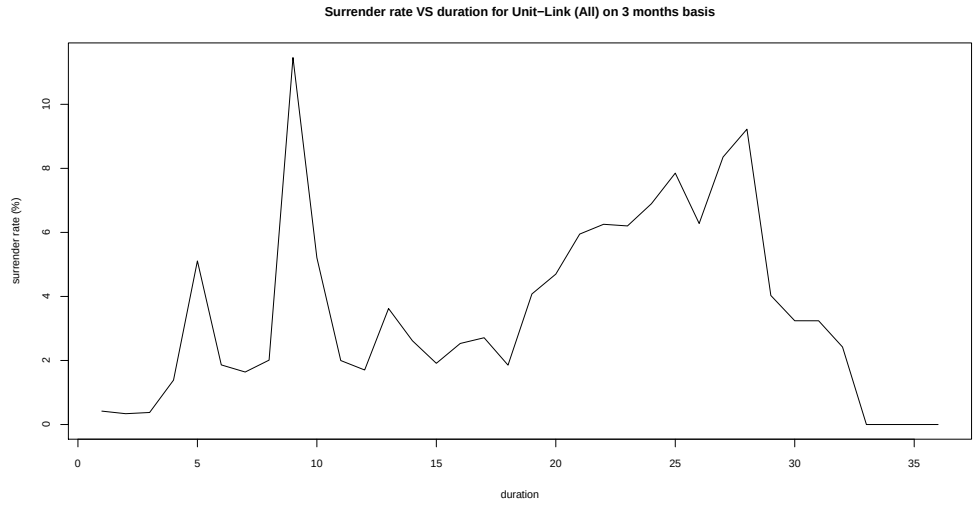
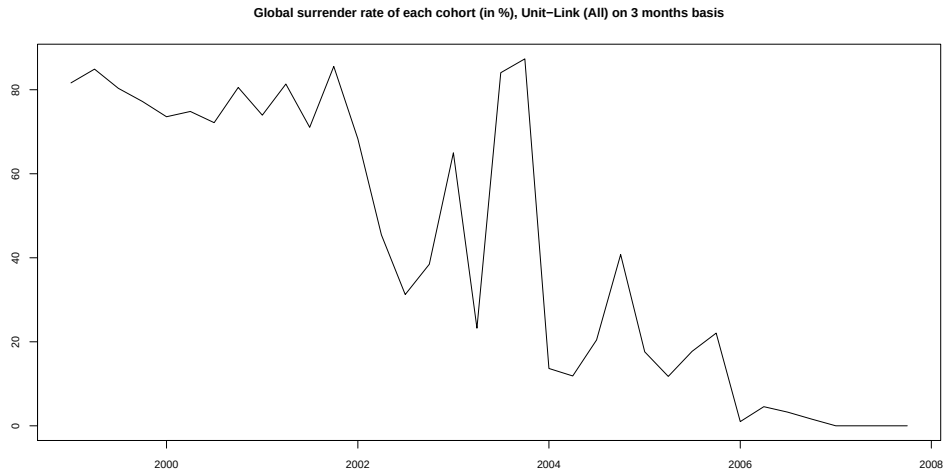


FIGURE 5.8 – Rachat par ancienneté de contrat (en trimestre) pour les produits UC.

FIGURE 5.9 – Pourcentage global de rachat par cohorte pour les produits UC.



Profil des rachats par ancienneté de contrat Le graphe 5.8 ne montre aucune relation monotone entre le taux de rachat et l'ancienneté de contrat, et met en évidence un fort pic de rachat à la fin de la deuxième année de contrat (8 et 9 trimestres d'ancienneté) mais ce comportement semble marginal et n'a pas d'explication rationnelle (de type frais spécifique pour un rachat à tel ou tel moment). L'allure de cette courbe nous fait pencher encore une fois pour une catégorisation de la variable "ancienneté" dans la modélisation, qui a l'avantage de mieux rendre compte de cette forme non-monotone mais qui a l'inconvénient d'augmenter le nombre de paramètres à estimer.

Taux de rachat par cohorte Le taux de rachat global par cohorte du graphique 5.9 prouve une forte hétérogénéité des comportements de rachat en fonction de la date

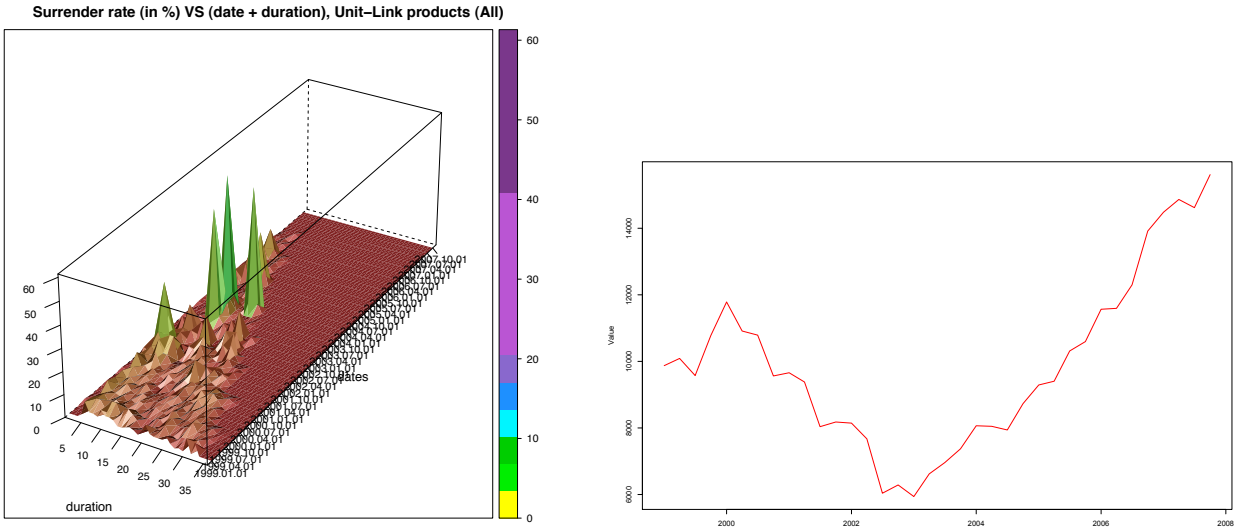


FIGURE 5.10 – A gauche : profil 3D du taux de rachat des UC par date et par ancienneté de contrat (par trimestre). A droite : évolution trimestrielle en valeur de l'indice boursier espagnol Ibex 35.

d'entrée en portefeuille. Le fait que les vieilles cohortes aient un taux élevé (environ 80 %) est tout à fait normal compte tenu de leur ancienneté, mais le pic de taux à presque 90 % pour les cohortes de fin 2003 paraît étonnant. Le couplage des graphes 5.8 et 5.9 pourrait d'ailleurs expliquer le pic observé fin 2005 sur le graphique 5.7 si le pic de rachat autour de 2 ans d'ancienneté de contrat est dû aux cohortes entrées fin 2003. Ce pic ne semble pas dû aux marchés financiers (car ceux-ci sont en nette hausse en 2005), mais plutôt à une nouvelle vague de produits UC que les agents d'AXA ont déployé en faisant racheter par là-même leur ancien contrat UC aux assurés.

Taux de rachat par date et par ancienneté de contrat Nous retrouvons dans le graphe 5.10 la confirmation que ce sont bien les cohortes qui ont souscrit fin 2003 qui rachètent fin 2005, mais pas pour des conditions de marché défavorables. Le pic de rachat en 2002 semble par contre correspondre à des comportements de rachat rationnels (dûs à une baisse du marché), ce qui induit une forte hétérogénéité pour la modélisation car les sources de rachat varient en plus de la volatilité des marchés.

5.2.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre L'erreur de classification (sur l'échantillon de validation) s'élève à 4,9 %, dont des indices de performance aux résultats exceptionnels de 92,6 % (pour la sensibilité) et de 96,5 % (pour la spécificité). Les comportements de rachat semblent donc très bien prédits par le modèle grâce aux variables dont nous disposons. Nous verrons dans la modélisation par régression logistique dynamique que les facteurs conjoncturels font voler en éclat ce constat.

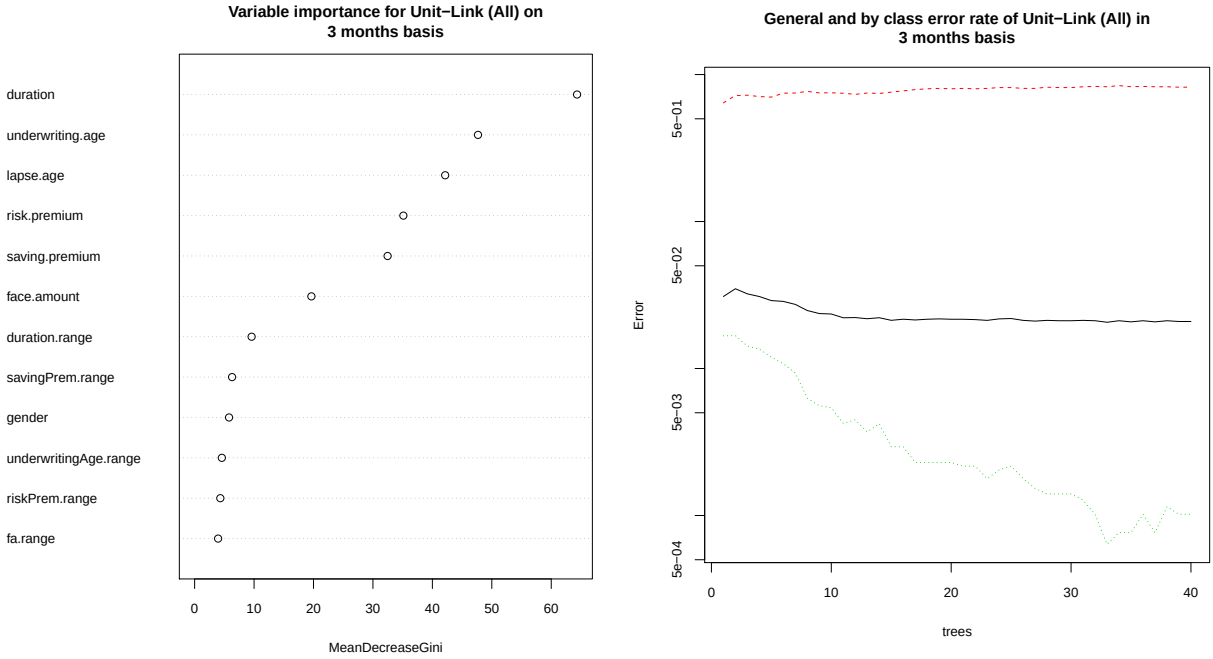


FIGURE 5.11 – Importance des variables explicatives, produit UC.

	Rachats non-observés	Rachat observés
Rachats non-prédits	4541	164
Rachats prédits	193	2410

Importance des variables explicatives Le classement de l'importance des variables explicatives disponible en figure 5.11 fait la part belle à l'ancienneté de contrat, l'âge de souscription, l'âge du rachat (corrélé à l'âge de souscription et l'ancienneté), la prime de risque et la prime d'épargne. Nous ne souhaitons sélectionner que les deux ou trois variables les plus importantes dans la modélisation pour minimiser la complexité du modèle final, ce qui donne (en considérant les variables catégorisées) l'ancienneté de contrat et la prime de risque (nous aurions pu considérer la tranche d'âge mais les statistiques desciptives nous montrent qu'en réalité cette variable n'est pas si discriminante).

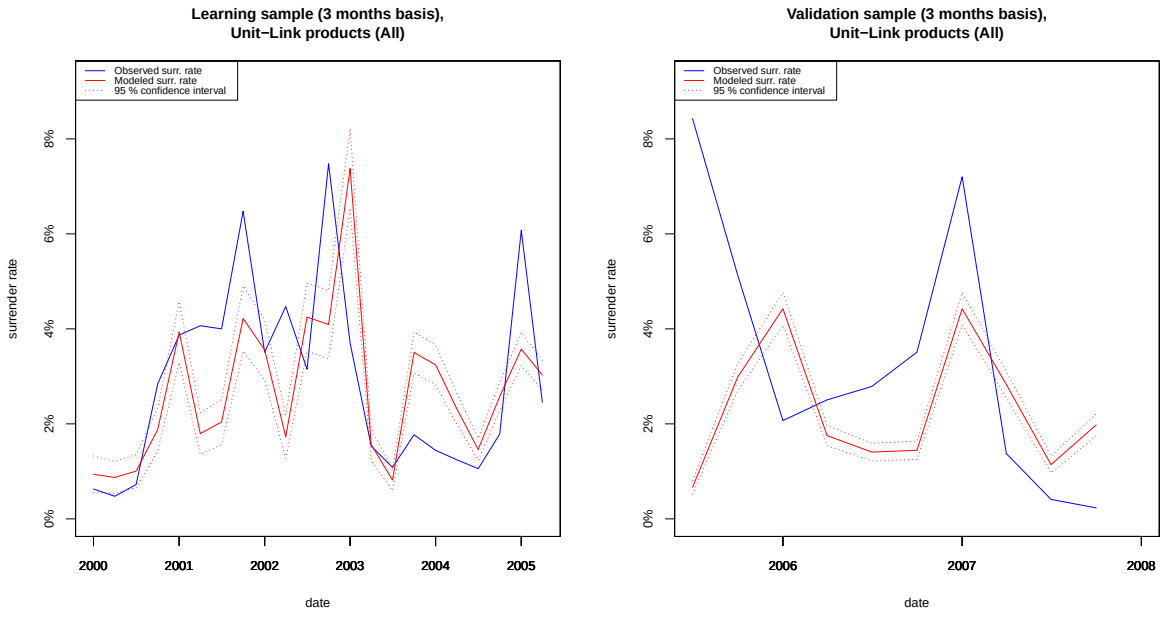
5.2.3 Modélisation et prévisions par mélange de GLM

Le contexte des produits en Unités de Compte est particulier puisque ceux-ci sont indexés sur les marchés financiers. Nous connaissons la volatilité du marché, qui a ainsi un impact direct sur la volatilité du taux de rachat lui-même. Mêlée aux effets "cohortes", une hétérogénéité très forte apparaît pour ce type de produit, pour lequel les comportements de rachat sont donc très difficilement prévisibles comme illustré par le graphe 5.12. C'est certainement dans ce contexte que la modélisation mélange a le plus d'apport. La sensibilité des assurés aux mouvements des marchés est évidemment très hétérogène. Le graphique 5.13 permet de constater que les principaux effets sont bien captés par le modèle, en bonne proportion et dans le bon sens. Le taux de rachat observé appartient à l'intervalle de confiance des prévisions sur toute la période (excepté fin 2005 et fin 2006), et ce malgré notre méthode de validation temporelle. La différence de la quantification de l'effet des marchés financiers entre la modélisation classique et la modélisation mélange est très importante, tant en termes de sens de l'impact que d'intensité.

Impact des variables explicatives par les mélanges de Logit L'annexe E.2.1 donne l'estimation des coefficients de régression du modèle mélange. Nous adoptons toujours la même méthode de choix d'estimation des variables (structurelles → coeff. identiques, conjoncturelles → coeff. variables) en espérant que les résultats soient probants. Détaillons maintenant les impacts des facteurs de risque :

- effets *structurels* : identiques à tout le monde. Pour la saisonnalité, les conclusions sont ressemblantes avec celles des contrats de pure épargne (cycle du marché de vente) : l'été est une période où très peu de rachats sont observés. L'effet de l'ancienneté du contrat est nettement moins évident comme le graphe 5.8 l'avait laissé présager. La richesse de l'assuré semble peu jouer, même si les personnes les plus riches ont l'air de racheter davantage (peut-être conseillées par un agent qui gère leur fortune).
- *effets conjoncturels* : une très grande majorité d'assurés (sauf ceux de la composante 4) rachète plus lorsque l'indice boursier plonge, mais il existe un petit groupe de personnes pour qui ce n'est pas le cas du tout (valeur positive du coefficient élevée). D'un trimestre à l'autre, la probabilité de rachat individuelle des assurés

FIGURE 5.12 – Modélisation et prévision du taux de rachat des produits UC par régression logistique dynamique.



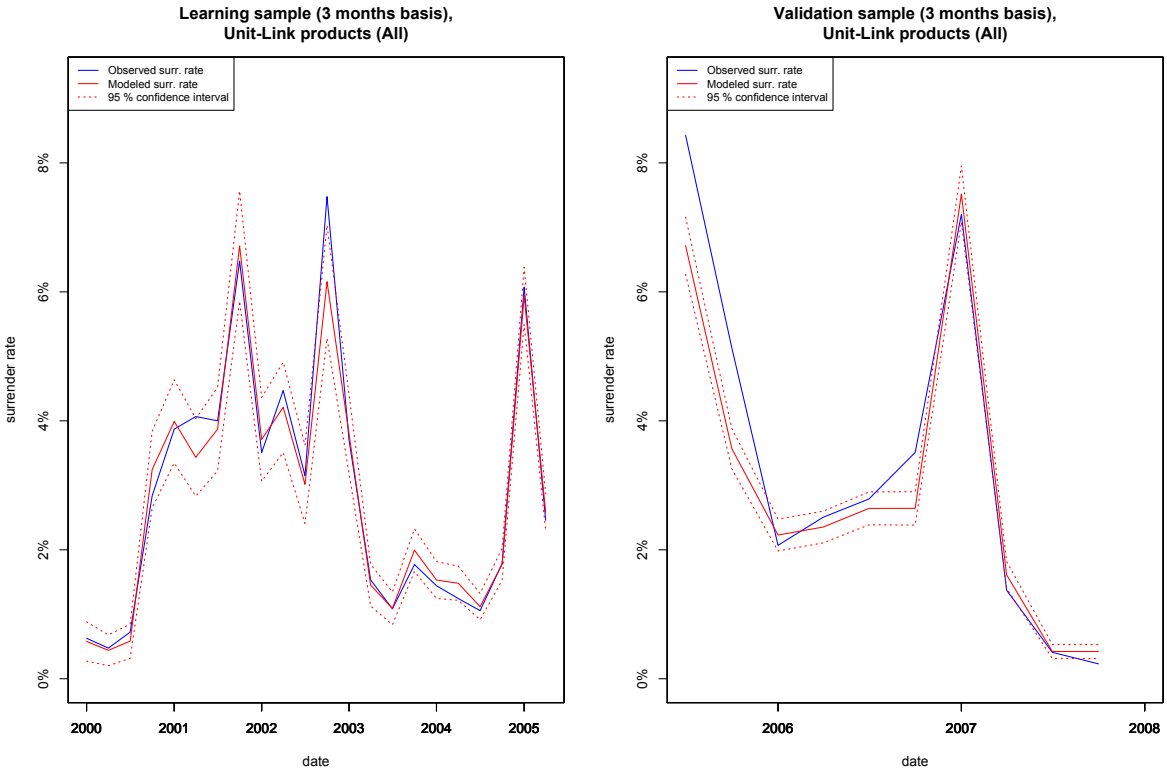


FIGURE 5.13 – Modélisation et prévision du taux de rachat par mélange de Logit, produits UC.

appartenant à un groupe (composante) donné change en fonction de la valeur de l'Ibex 35, créant ainsi une corrélation positive entre les personnes de ce groupe.

Les mêmes constatations peuvent être formulées concernant l'irrationnalité et l'hétérogénéité des réactions des assurés, même si la proportion d'agents irrationels paraît ici très limitée. Il semblerait que la rationalité des agents soit plus forte pour ce type de produit.

5.3 Les contrats liés au indices boursiers (Index-Link)

Ce type de contrat est très semblable aux contrats en UC. La différence réside dans le support d'investissement qui est ici plus ciblé car il s'agit uniquement d'indices boursiers. Les variables assurés et contrat dont nous disposons sont identiques aux types de produit précédent (issu de la même base données) et nous étudions donc ces contrats sur la période 1/1/2000-31/12/2007. Nous devrions logiquement obtenir des résultats en ligne avec l'étude des produits en UC.

5.3.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille La première remarque que nous pouvons formuler avec le graphe 5.14 est que le taux de rachat est globalement très faible, écrasé par le taux de chute dans le graphe d'origine (c'est la raison pour laquelle nous ne traçons pas le taux de chute ici). La deuxième constatation est que l'exposition d'AXA Seguros à ce type de produit est moindre, en très forte baisse depuis 2005, et que nous n'observons pas clairement de saisonnalité. Le manque de diversification du support d'investissement joue certainement un rôle dans cette statistique, les vendeurs et les souscripteurs connaissant de plus en plus l'importance de cette diversification pour diminuer le risque global, d'où une formule peu attractive. Il n'y a effectivement plus de nouvelle souscription sur ce type de produits depuis début 2005.

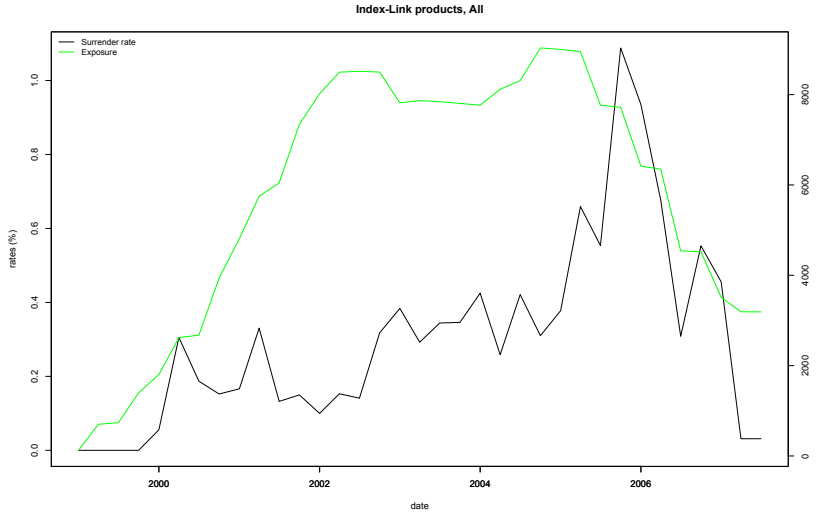
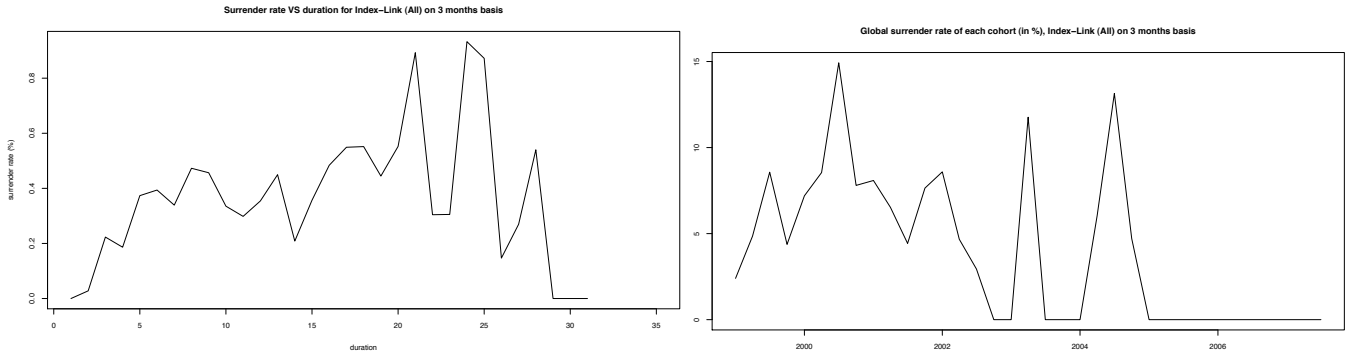


FIGURE 5.14 – Exposition et taux de rachat trimestriel du portefeuille de produits Index-Link.

FIGURE 5.15 – A gauche : rachat par ancienneté de contrat (en trimestre). A droite : Pourcentage global de rachat par cohorte. Produits Index-Link.



Profil des rachats par ancienneté de contrat et taux de rachat par cohorte En un certain sens, le profil des rachats en fonction de l'ancienneté des contrats du graphe 5.15 rappelle celui constaté sur les produits en UC. Il en est de même concernant les taux de rachat globaux par cohorte (c'est pourquoi nous regroupons ici ces deux graphiques). Une forme erratique, imprévisible, non-monotone. La différence majeure concerne le comportement des cohortes qui semble davantage directement lié à l'indice Ibex 35, qui rappelons le s'effondre entre 2000 et 2002, provoquant un niveau moyen de rachat des cohortes supérieur visible sur cette période.

Taux de rachat par date et par ancienneté de contrat La mise en évidence d'une forte hétérogénéité par le graphique 5.16 vient confirmer l'ensemble des observations

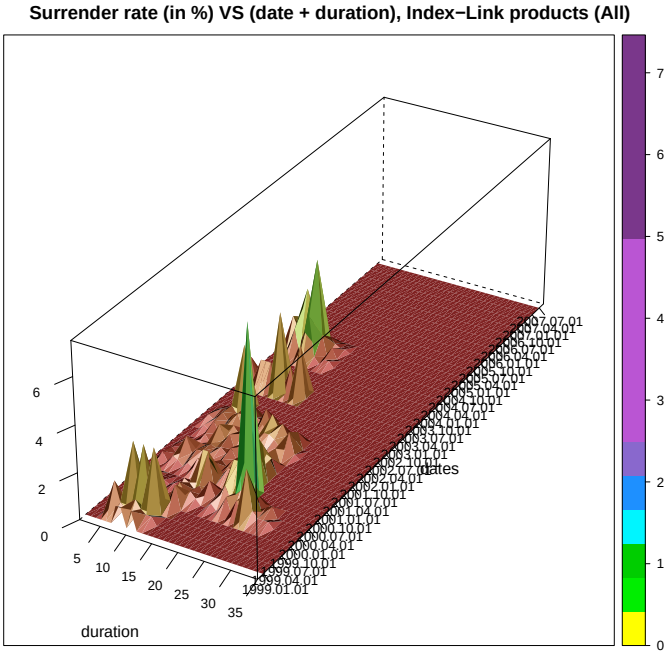


FIGURE 5.16 – Profil 3D du taux de rachat par date et par ancienneté de contrat, Index-Link.

faites précédemment. Il ne se dégage pas de profil précis en fonction de l'ancienneté de contrat, mais la date calendaire (ici entre 2000 et 2002) et donc le contexte économique joue clairement un rôle. La vague de rachat de fin 2005 était déjà observée sur les produits en UC, et ne correspond toujours pas à la chute de l'indice boursier. Nous évoquons des politiques de vente pour expliquer ce fort pic (il semblerait qu'il y ait peut-être eu des problèmes avec les réseaux de distribution à cette période mais l'information n'est pas directement disponible dans la base de données).

5.3.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre Le classifieur par forêts aléatoires se trompe rarement dans la prévision des rachats lorsque ceux-ci sont effectivement observés (erreur que nous cherchons à minimiser car la plus risquée pour nous), donnant une spécificité rassurante de 99 %. La sensibilité vaut ici 72 % et l'erreur globale de classification est égale à 3,6 %. Le classifieur est très précis sur l'étude statique.

	Rachats non-observés	Rachat observés
Rachats non-prédits	6770	70
Rachats prédits	204	526

Importance des variables explicatives Nous avons choisi cette fois de montrer le classifieur sous forme d'arbre (figure 5.17) de classification par la méthode échantillon témoin-échantillon de validation. Nous avons sensiblement le même classement que pour les produits en UC, avec une certaine importance de l'âge de souscription. L'intérêt de cette représentation est la définition de seuils précis, utiles dans un processus de segmentation. Nous retenons donc l'ancienneté du contrat et l'âge de souscription.

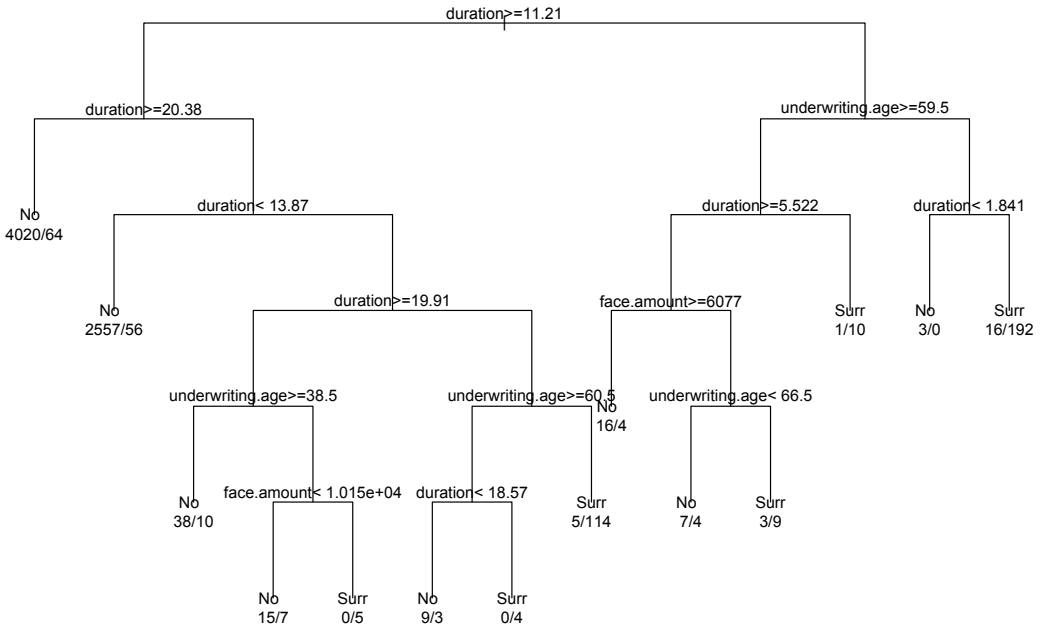


FIGURE 5.17 – Arbre de classification, donnant l'importance des variables explicatives des produits Index-Link en partant de la racine vers les feuilles.

5.3.3 Modélisation et prévisions par mélange de GLM

Comme pour les produits en UC, le modèle statistique de régression logistique dynamique est inefficace tant sur la période d'échantillonnage où il ne reflète pas les pics en début de période, que sur la période validation où le niveau de rachat n'est pas bien ajusté (cf graphe 5.18). La conclusion de ce constat est que certes les effets économiques sont modélisés, mais la calibration de ces effets n'est visiblement pas adéquate. A l'inverse, le graphique 5.19 est très satisfaisant, le taux de rachat observé aussi bien sur la période d'apprentissage que sur la période validation reste toujours dans l'intervalle de confiance du taux prédit. La prise en compte spécifique des variables explicatives dans le mélange permet d'arriver à ces résultats, avec toujours en idée de modéliser l'hétérogénéité par des coefficients de régression variables entre composantes pour les effets conjoncturels.

Impact des variables explicatives par les mélanges de Logit L'estimation des coefficients de régression du modèle mélange en annexe E.3.1 dévoile moins d'hétérogénéité que pour les produits en UC, peut-être à cause du fait que le support des produits soit relativement simple à interpréter (seul l'ibex 35 sert de valorisation au contrat). Cela rend la compréhension du produit et l'interprétation des résultats du contrat plus simples pour l'assuré, contrairement à un produit qui serait indexé sur plusieurs supports et dont l'assuré aurait du mal à savoir de manière globale la valeur. Nous adoptons toujours la même méthode de choix d'estimation des variables (structurelles → coeff. identiques, conjoncturelles → coeff. variables). Les impacts des facteurs de risque sont les suivants :

– effets *structurels* : identiques à toutes les composantes. Pas d'effet saisonnalité introduit sur ce type de produit, l'ancienneté du contrat joue toujours dans le même sens (les assurés rachètent globalement rapidement). Les hommes rachètent

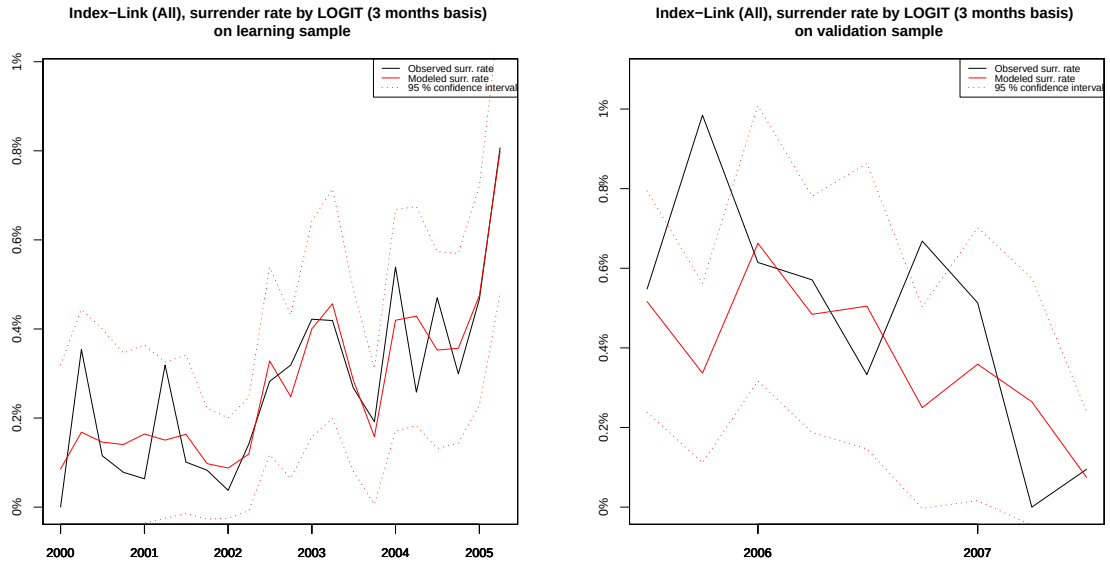
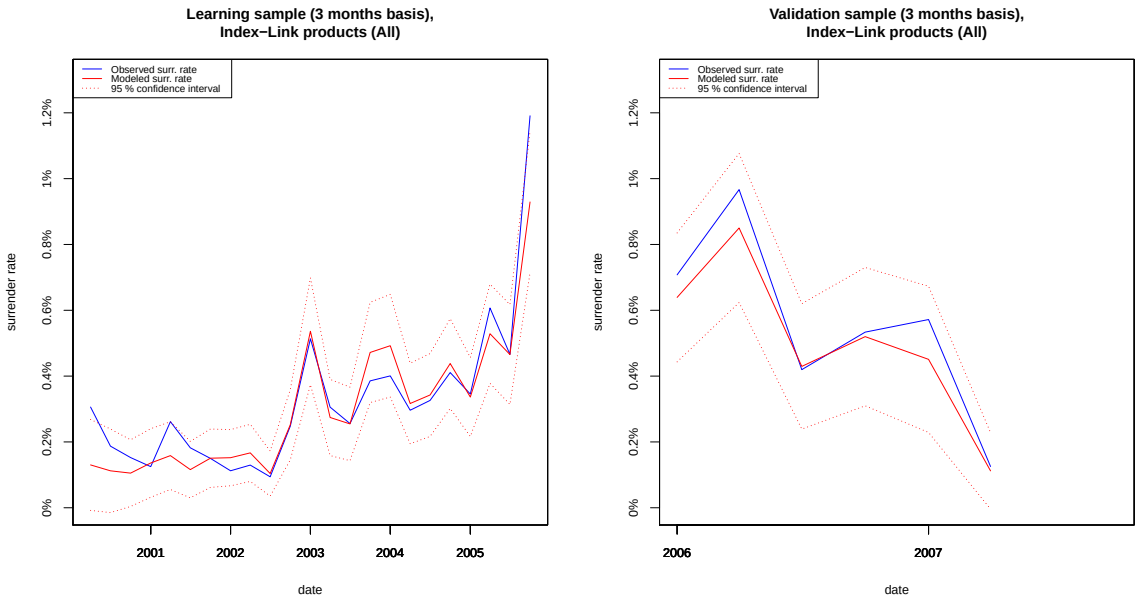


FIGURE 5.18 – Modélisation et prévision du taux de rachat des produits Index-Link par régression logistique dynamique.

FIGURE 5.19 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Index-Link.



plus que les femmes (effet ajouté car visible dans les statistiques descriptives mais non retranscrit par les arbres), et les personnes âgées semblent racheter moins souvent que les autres.

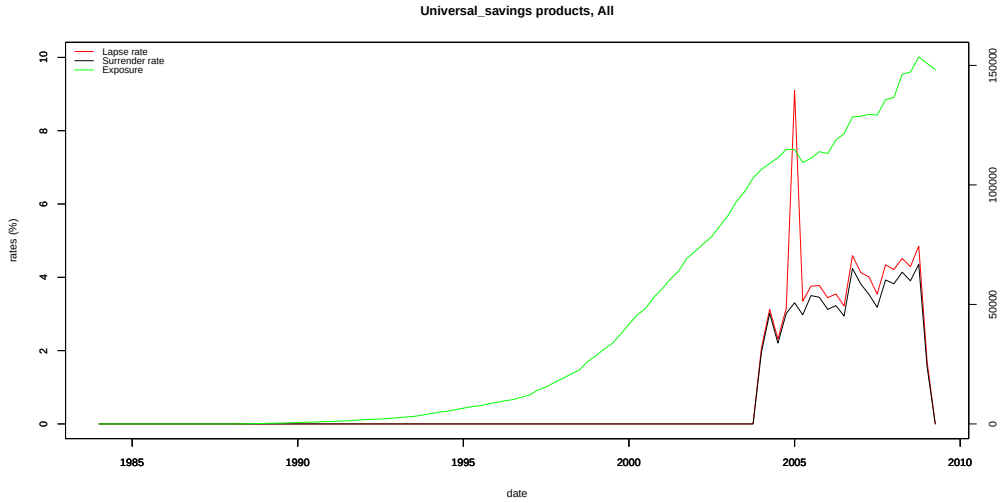
– effets *conjoncturels* : le critère BIC sélectionne un mélange à seulement deux composantes. L'hétérogénéité est donc moins grande à priori, d'ailleurs l'ensemble des assurés réagit aux mouvements de l'Ibex 35 dans le même sens (seule la sensibilité à ces mouvements est plus ou moins grande). La probabilité de rachat individuelle des assurés augmente exponentiellement en fonction de l'évolution de l'Ibex 35, avec la même amplitude pour tous les individus appartenant à la même composante (corrélation positive entre les comportements).

Nous remarquons que le calibrage du modèle pour cette famille de produit ne nécessite que deux composantes (annexe E.3.1), et que l'estimation des proportions du mélange semble robuste.

5.4 Famille "Universal savings"

Ces contrats d'épargne à taux garanti offrent des garanties prévoyance supplémentaires. En général, il s'agit de garantie classique contre le décès de l'assuré, mais certains assureurs proposent des options supplémentaires (appelées "rider") comme par exemple des garanties d'incapacité ou invalidité. Dans leur fonctionnement les Universal Savings se rapprochent des Ahorro, mais la composante Prévoyance vient sûrement modifier l'usage qui en est fait par l'assuré. La période d'observation va de 1985 à fin 2009, et les variables explicatives dont nous disposons sont le numéro du produit, la date d'émission, la date de sortie et sa raison (si sortie il y a), la date de naissance de l'assuré, son sexe, sa richesse, la fréquence de la prime, la prime de risque, la prime d'épargne et le réseau de distribution. Dommage que l'on ne dispose pas de l'information sur le type de prime

FIGURE 5.20 – Exposition et taux de rachat trimestriel du portefeuille des produit Universal Savings.



(nivelée ou non) qui doit être un facteur explicatif important (effet psychologique).

5.4.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Comme nous pouvons le constater dans le graphique 5.20, les chutes (dont les rachats) n'ont malheureusement été enregistrées dans la base de données qu'à partir de Janvier 2004. L'étude de cette grande famille de produit est intéressante car l'exposition est très importante (elle va jusqu'à 150 000 contrats simultanément en portefeuille). Là encore, la crise financière semble avoir joué un rôle étant donné la chute constatée à partir de 2008. Hormis cette chute, la tendance des rachats était à une croissance légère et continue, peu volatile et présentant une certaine périodicité.

Profil des rachats par ancienneté de contrat et taux de rachat par cohorte En exceptant les premières années pour lesquelles nous retrouvons le même type de

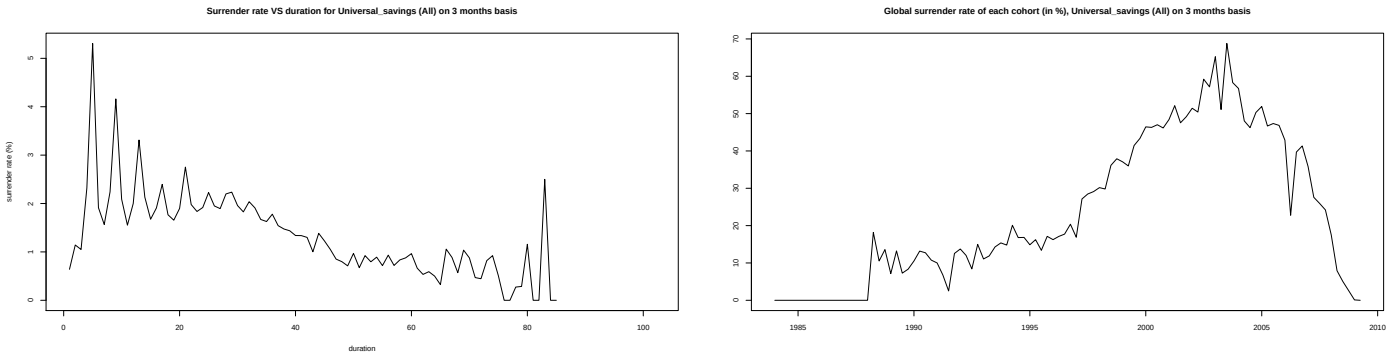


FIGURE 5.21 – A gauche : rachat par ancienneté de contrat (en trimestre). A droite : Pourcentage global de rachat par cohorte. Produits Universal Savings.

profil que pour les produits de pure épargne, la décroissance est plutôt linéaire ensuite, avant de se terminer par un pic dont il semblerait logiquement qu'il soit lié à quelques cohortes particulières. L'autre différence concerne le taux de rachat global par cohorte : bizarrement les cohortes les plus jeunes ont déjà plus racheté que les anciennes (croissance linéaire), ce qui traduit clairement une évolution des moeurs des assurés sur les comportements de rachat. Les enseignements que nous pouvons tirer des graphiques de la figure 5.21 sont la nécessité une fois de plus de rendre la variable "ancienneté" catégorielle, et une hétérogénéité entre générations créée par un facteur non-observable. Nous n'affichons pas le taux de rachat par date et par ancienneté de contrat car les conclusions sont identiques aux précédentes.

5.4.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre Le processus de classification (sur l'échantillon de validation) résumé dans le tableau ci-dessous donne un taux d'erreur de 4.1 %. L'erreur liée à la spécificité (82.8 % de bonnes prévisions) est un peu plus grande que celle de la sensibilité (98.1 %), mais globalement les forêts aléatoires ont un bon pouvoir de prévision sur ce jeu de données.

	Rachats non-observés	Rachat observés
Rachats non-prédits	11972	2480
Rachats prédits	1700	85840

Importance des variables explicatives Il faut noter que pour l'une des premières fois ce n'est pas l'ancienneté du contrat qui apparait comme le facteur de risque le plus discriminant par rapport au comportement de rachat. La richesse de l'assuré est clef dans

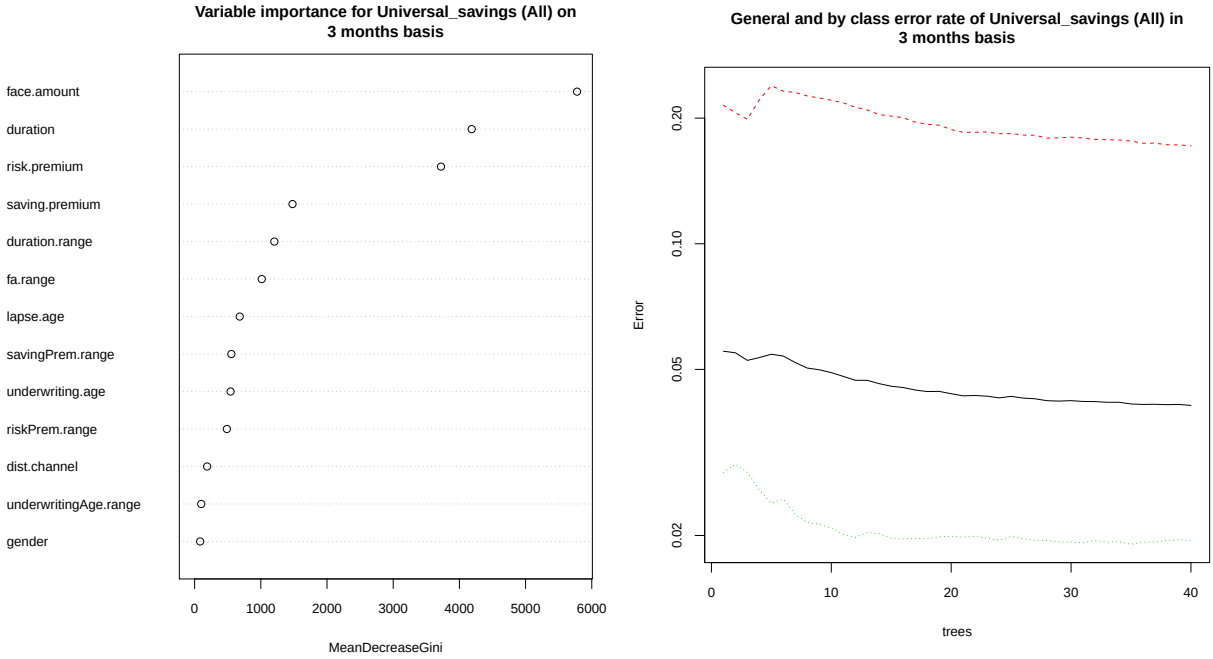


FIGURE 5.22 – Importance des variables explicatives, produit Universal Savings.

le processus de décision au vu des résultats de la figure 5.22, suivie de l'ancienneté de contrat et de la prime de risque et d'épargne. Les niveaux de ces différentes primes sont corrélés à la richesse de l'assuré : nous prenons donc uniquement cette dernière variable dans la modélisation. Malgré leur impact relativement faible au vu de ce graphique, le réseau de distribution et l'âge de souscription seront introduits dans la modélisation de la décision de rachat pour gagner en précision.

5.4.3 Modélisation et prévisions par mélange de GLM

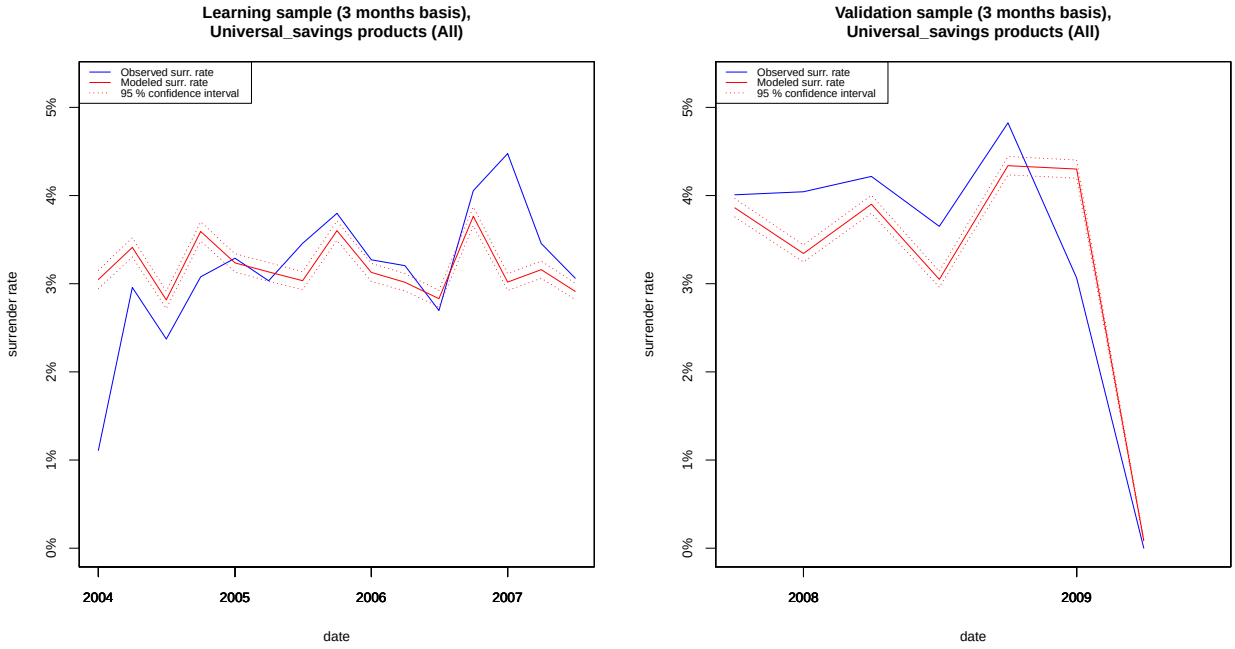
La chute du taux de rachat fin 2008 est relativement bien capté dans le modèle simple de régression logistique dynamique, ce qui veut dire que l'impact du facteur provoquant cette chute a été bien quantifié dans la procédure. Par contre, les niveaux et variations du taux de rachat sur la période d'apprentissage sont mal ajustés par le modèle. Les effets semblent pourtant être modélisés dans le bons sens (les variations se font dans la même direction sur la courbe 5.23), au détriment de l'amplitude qui manque de précision. Le mélange de régressions logistiques permet non seulement de capter les bons effets mais surtout de rendre compte de l'hétérogénéité présente dans les données, en capturant parfaitement le comportement des groupes de personnes et en restituant une modélisation très précise du taux de rachat, aggrégation de l'ensemble des décisions individuelles (graphe 5.24). De plus, le taux observé restant dans l'intervalle de confiance des prévisions (malgré son étroitesse) tout au long de la période d'étude confirme la parfaite adéquation du modèle.

Impact des variables explicatives par les mélanges de Logit La découverte que nous avons faite pour cette famille de produit est étonnante : il ne semble pas nécessaire de prendre en compte le contexte économique pour avoir une modélisation précise. Les graphiques ci-dessus sont issus de modélisations logistiques à une ou plusieurs composantes, mais pour lesquelles aucune variable de type taux long-terme ou Ibex 35 n'ont été introduites. La caractéristique du type de produit, qui accorde une plus grande importance aux garanties de prévoyance (liées aux risques de la vie), change visiblement la manière qu'ont les assurés de prendre leur décision. La couverture prévoyance semble prendre le dessus sur l'environnement économique, prouvant que le produit est perçu davantage comme un produit de prévoyance que comme un produit d'épargne.

L'estimation des coefficients de régression du modèle mélange en annexe E.4.1 confirme l'importance de la variable "richesse" dans le processus de décision. L'hétérogénéité des comportements est capté principalement grâce à cette variable (le risque de base donné par l'intercept est presque le même pour toutes les composantes). Les impacts des facteurs de risque sont les suivants :

- effets *structurels* : faible effet de saisonnalité avec une augmentation des rachats en fin d'année civile, l'effet de l'ancienneté du contrat est sensiblement différent. Les assurés semblent racheter leur contrat en moyenne plus tard (ancienneté moyenne). Les personnes âgées rachètent moins (besoin accrue de se couvrir contre les risques vie), et le réseau de distribution est discriminant : un suivi du contrat plus personnalisé conduit à une diminution de rachats.
- effets *conjoncturels* : ce sont les assurés de richesse intermédiaire qui rachètent le plus. Vu les poids des composantes (annexe E.9), la plupart des assurés adopte ce comportement (composantes 3 et 4). Cependant, d'autres (une faible minorité) ont un comportement différent et rachète davantage malgré leur grande richesse.

FIGURE 5.23 – Modélisation et prévision du taux de rachat des produits Universal Savings par régression logistique dynamique.



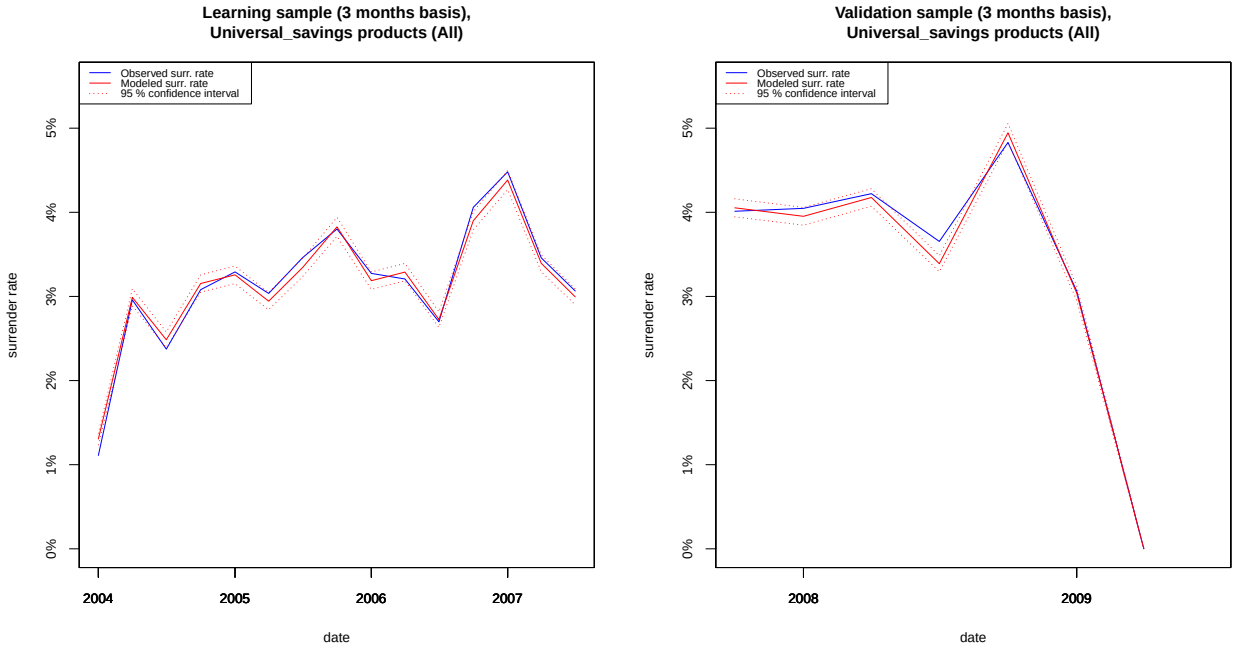


FIGURE 5.24 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Universal Savings.

Le mélange comporte cinq composantes, en proportion respective de la composante 1 à 5 : 13 %, 10 %, 30 %, 35 % et 12 %). L'estimation robuste de ces poids confirment que chaque composante joue un rôle de capture d'un type de comportement, permettant au modèle de reéquilibrer les effets par période en fonction de la composition du portefeuille.

5.5 Les contrats à taux garanti : les "Pure savings".

Cette gamme de contrat peut s'apparenter complètement aux contrats de type "Ahorro". La dénomination diffère car la base de données que nous utilisons pour les étudier est différente de celle des "Ahorro", plus complète et couvrant une plus grande période : de 1967 à fin 2009 (mais comme précédemment, les rachats n'ont été répertoriés qu'à partir du 1/1/2004). Les informations disponibles sont identiques à celles concernant les produits Universal Savings car la base de données d'origine est la même.

5.5.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Le constat du graphique 5.25 est que le taux de rachat semble avoir une saisonnalité (périodicité), avec un taux de rachat qui semble augmenter et baisser à des intervalles de temps réguliers. Sur la fin de la période d'observation, nous remarquons un comportement anormal de la courbe avec un forte hausse suivie d'une baisse brusque et importante. Nous allons voir dans la suite s'il est possible d'expliquer ces mouvements. L'exposition sur cette ligne de produit peut aller jusqu'à 80 000 contrats, ce qui laisse présager des résultats statistiques robustes (qu'ils soient bons ou mauvais).

Profil des rachats par ancienneté de contrat Le profil des rachats en fonction de l'ancienneté des contrats donné par le graphe 5.27 semble indiquer une forme en cloche, avec une majorité des assurés qui rachète en moyenne vers le 50ème trimestre (13ème année). Cette forme n'a d'explication ni par les conditions des contrats quant

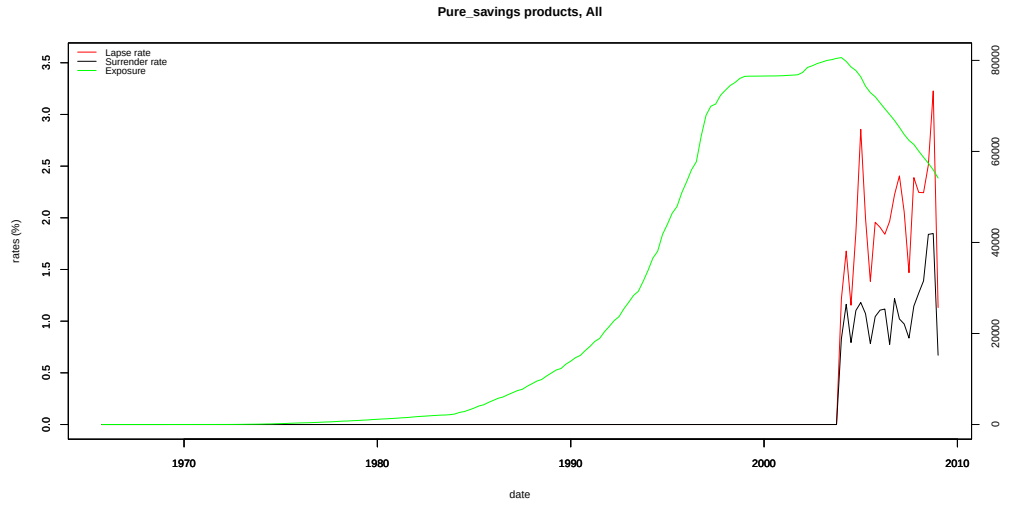
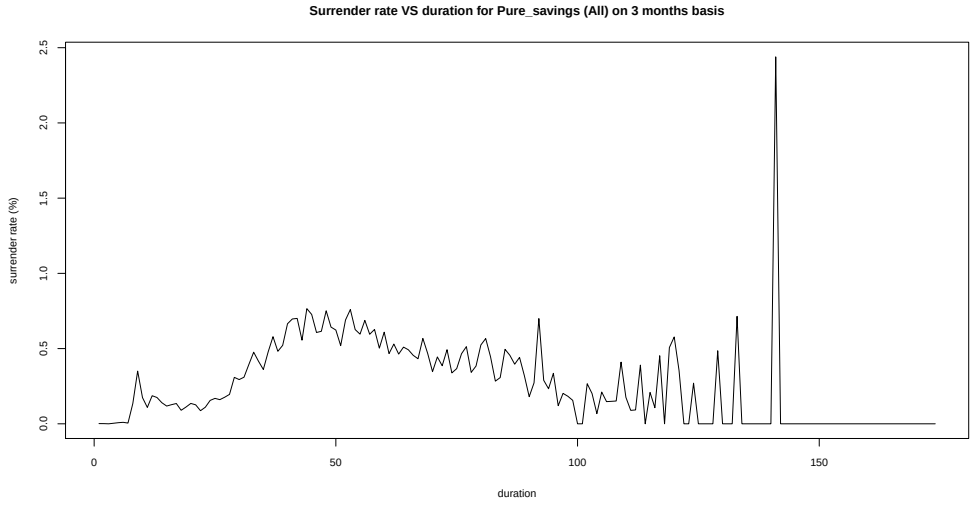


FIGURE 5.25 – Exposition et taux de rachat trimestriel du portefeuille de produits Pure Savings.

FIGURE 5.26 – Rachat par ancienneté de contrat (en trimestre) pour les produits Pure Savings.



aux rachats, ni par la fiscalité. Cela reste donc une donnée statistique dont il faudra tenir compte lors de l'introduction de la variable "ancienneté" dans la modélisation (sous forme catégorielle de préférence donc).

Taux de rachat par cohorte et nouvelles affaires Nous avons décidé d'afficher l'évolution des nouvelles affaires sur ce produit car un phénomène peu courant est à l'origine de l'"hétérogénéité" constatée sur la figure 5.27 (partie gauche). En fait il n'y a pratiquement plus de nouvelles affaires souscrites fin 1999-début 2000 (entre 1 et 15 contrats par trimestre !), ce qui fait mécaniquement chuter le taux de rachat de ces cohortes lorsque le(s) seul(s) assuré(s) ayant souscrit n'a(ont) pas racheté. Ce phénomène est marginal et ne doit donc pas être interprêté comme des changements de comportements (hétérogènes), de même que le pic de la cohorte fin 2008 (1 seul contrat avait été souscrit et il a été racheté). Comme pour les produits Universal Savings, nous n'affichons pas le graphique 3D du taux de rachat en fonction de l'ancienneté du contrat et de la date car il n'amène pas d'information intéressante supplémentaire.

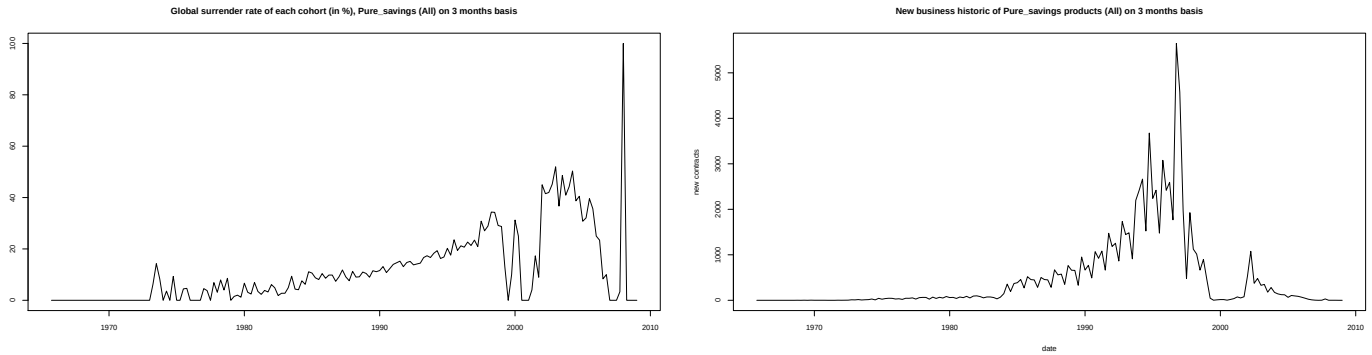


FIGURE 5.27 – A gauche : taux de rachat global par cohorte. A droite : nouvelles affaires par trimestre. Produits Pure Savings.

5.5.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre Les critères de performance de la classification des comportements de rachat sur les produits Pure Savings (échantillon de validation) sont quasiment similaires : la spécificité vaut 89.7 % tandis que la sensibilité vaut 88.7 %, pour un taux d'erreur global de mauvaise classification de 10.8 %. La méthode CART apparaît encore comme une bonne alternative de modèle de classification, malgré un taux d'erreur en hausse comparé aux précédentes applications.

	Rachats non-observés	Rachat observés
Rachats non-prédits	11046	1411
Rachats prédits	1602	13944

Importance des variables explicatives Au vu de la figure 5.28, les variables explicatives les plus discriminantes sont dans l'ordre décroissant : l'ancienneté de contrat, l'âge au moment du rachat, la prime d'épargne, l'âge de souscription, la richesse... Pour rester en ligne avec les hypothèses de modélisation (variables indépendantes en théorie) et les modélisations des autres produits, nous considérons la saisonnalité, l'ancienneté de contrat et l'âge de souscription (corrélé à l'âge au moment rachat). Ces trois facteurs explicatifs devraient suffire à effectuer nos prévisions de taux.

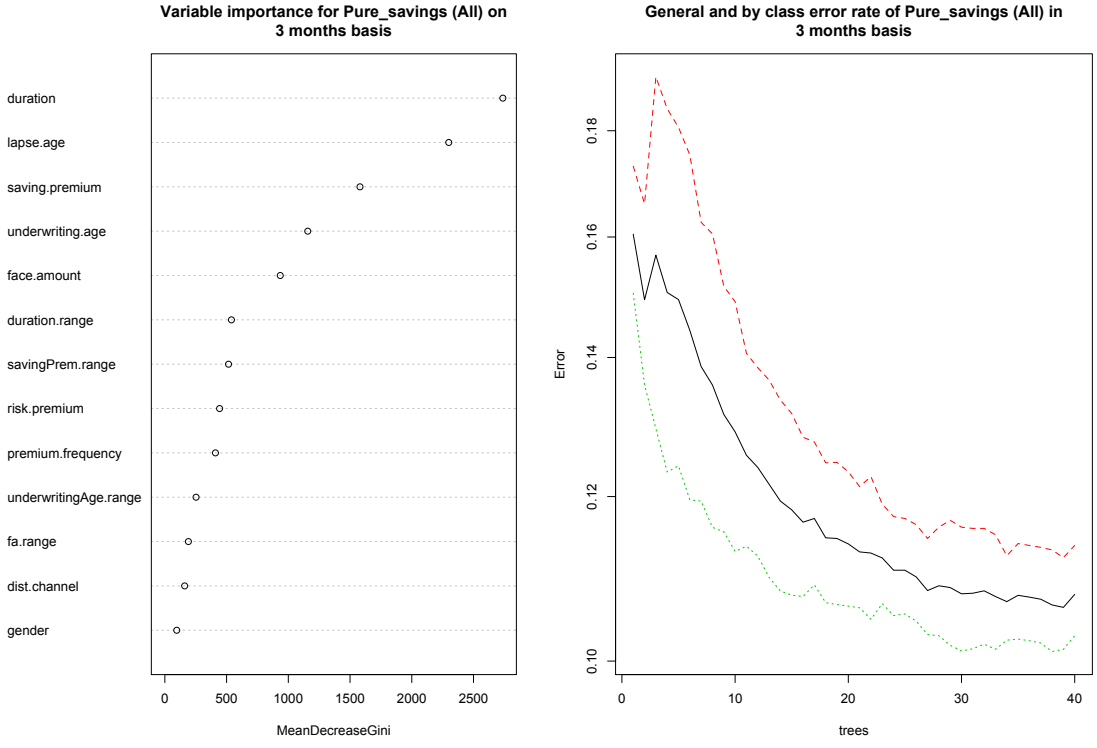


FIGURE 5.28 – Importance des variables explicatives, produit Pure Savings.

5.5.3 Modélisation et prévisions par mélange de GLM

Nous avons vu que l'hétérogénéité des données n'était pas si grande que d'habitude sur cette ligne de produit. Cette remarque est quelque part validée par le modèle de régression logistique dynamique, qui ne s'en sort pas si mal en termes de modélisation et de prévision (graphe 5.29). Quelques ajustements (notamment sur la période de validation) seraient préférables mais la dynamique du taux de rachat est grosso modo reproduite par le modèle. Qu'en est-il par l'usage des mélanges ?

Les produits de type Pure Savings sont l'unique cas dans toutes nos données où l'apport de la modélisation mélange n'est pas forcément évident. Le graphique 5.30 relate l'évolution du taux de rachat observé et du taux de rachat prédit et semble montrer un meilleur ajustement du modèle, mais au prix d'une certaine complexification. En effet, nous pouvons remarquer que le niveau est mal ajusté en milieu d'année 2008 (nous sommes d'ailleurs assez loin de la réalité), ce qui laisse supposer que certains effets non-observables ont joué à ce moment là, mais ne sont pas captés par la modélisation.

Impact des variables explicatives par les mélanges de Logit Le "boxplot" de l'estimation des coefficients de régression du modèle mélange disponible en annexe E.5.1 implique quatre composantes pour le mélange. Les impacts des facteurs de risque sont les suivants :

- effets *structurels* : un effet de saisonnalité prononcé (forte baisse des rachats en été, de juillet à septembre), l'effet de l'ancienneté du contrat est un peu différent : les assurés dont l'ancienneté appartient à la tranche la plus basse sont plus susceptibles (fortement) de racheter alors que ceux de la deuxième tranche rachète très sensiblement plus. Plus leur âge augmente et moins les assurés rachètent leur contrat.
- effets *conjoncturels* : l'Ibex 35 et le taux 10Y jouent toujours dans le même sens

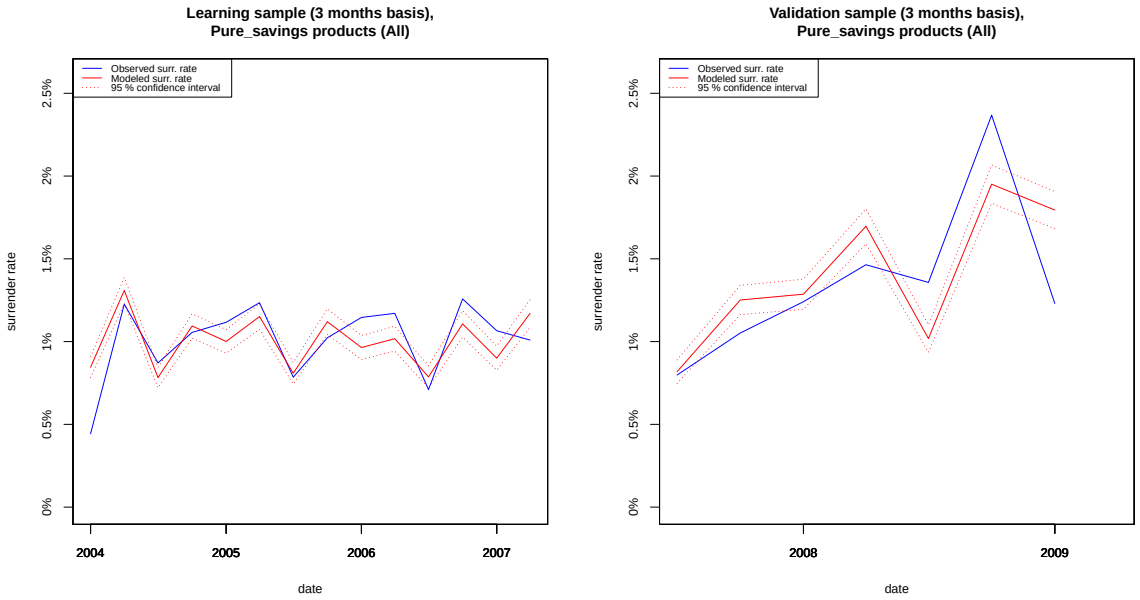
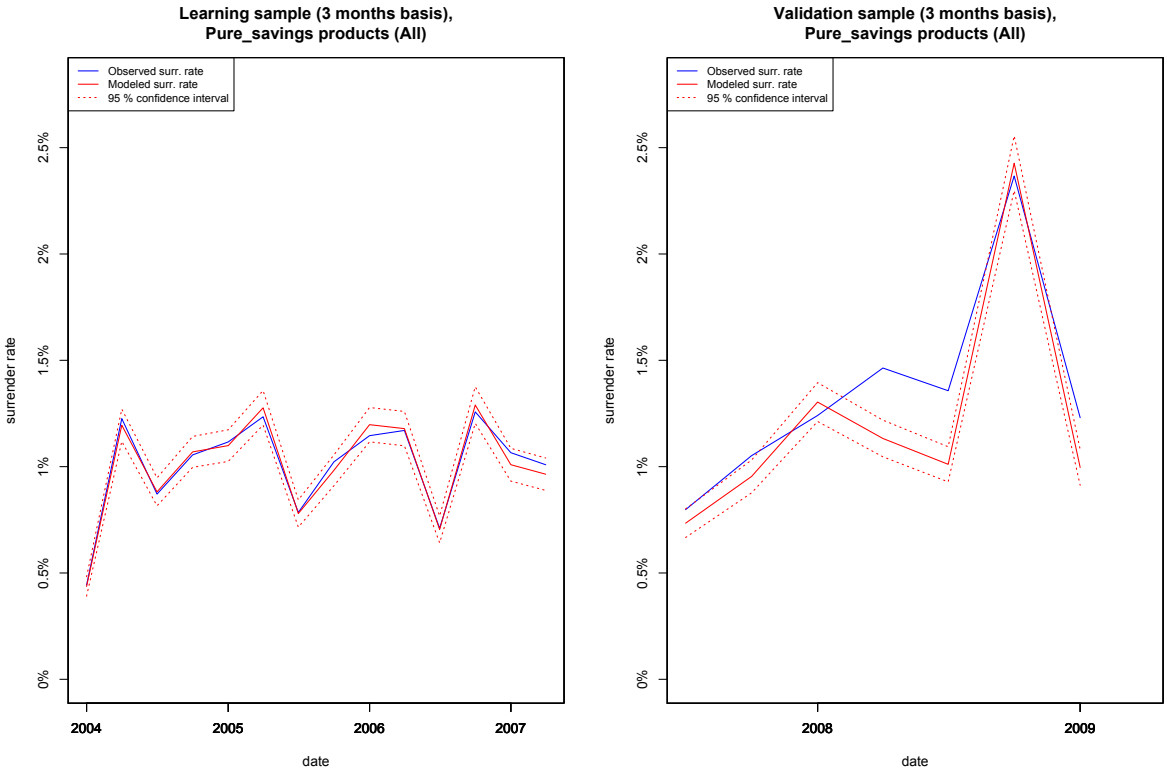


FIGURE 5.29 – Modélisation et prévision du taux de rachat des produits Pure Savings par régression logistique dynamique.

FIGURE 5.30 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Pure Savings.



pour ce type de produit. Leurs effets sont presque de même amplitude (avec un petit plus pour l'Ibex 35). Lorsque l'Ibex et le taux 10Y baissent, la probabilité individuelle de rachat augmente pour une grande majorité des assurés avec des groupes très sensibles (composantes 2 et 4) et d'autre moins sensible (composante 1). Au vu des poids des composantes (Annexe, figure E.11), un gros tiers des assurés (36 %) se comportent de manière contraire.

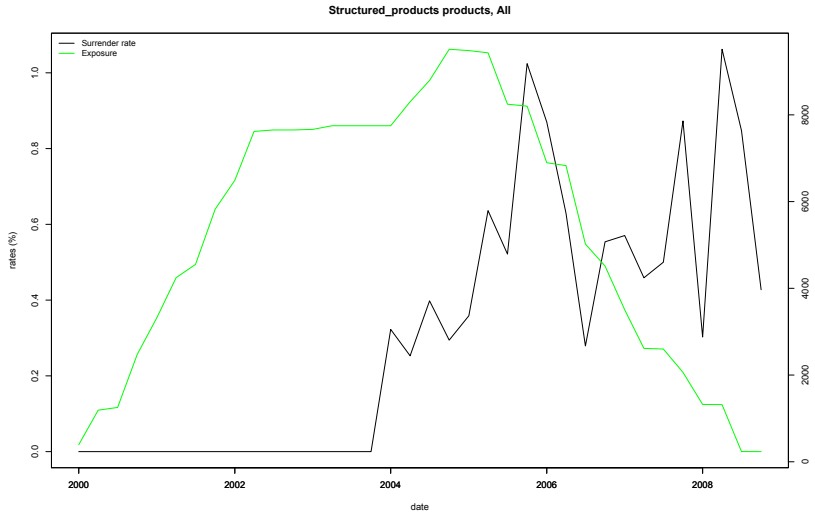
– *corrélation* : introduite via le contexte économique, toutes les personnes d'un groupe donné vont voir leur probabilité de rachat individuelle grimper ou chuter simultanément, faisant ainsi varier les prévisions de rachat au niveau global.

Le fait que les estimations des poids des composantes aient des écart-types importants (0 appartient à l'intervalle de confiance pour chaque estimation) doit partiellement être à l'origine du résultat mitigé que nous obtenons. Il faut toutefois avoir l'honnêteté de dire que nous avons cherché une meilleure modélisation sans pour autant la trouver, ceci est le signe que l'usage des mélanges ici n'est pas forcément pertinent.

5.6 Les produits structurés ou "Structured Products".

Il est relativement difficile de décrire les produits stucturés puisque par définition ils sont très variés. Ce sont en général des produits qui dépendent fortement de la performance des marchés financiers, dans le sillage des produits en UC ou des produits indexés sur les indices boursiers. Chaque produit a sa caractéristique, et ce sont en général les proportions d'investissement sur tel ou tel support qui varient entre les

FIGURE 5.31 – Exposition et taux de rachat trimestriel du portefeuille de produits Structurés.



produits. La période d'observation s'étend de début 2000 à fin 2009 (avec toujours les rachats enregistrés seulement depuis début 2004) et les informations disponibles sur le contrat et l'assuré sont identiques à celles des Pure Savings que nous venons d'étudier.

5.6.1 Analyse descriptive

Evolution de l'exposition et du taux de rachat du portefeuille Nous n'affichons pas le taux de chute car il écrase complètement le taux de rachat, signalant au passage que les comportements de rachat ne sont pas très nombreux sur cette ligne de produit (mais l'exposition est limitée par rapport aux produits dont nous venons de parler). Nous verrons par la suite que le taux de rachat présente en revanche une très forte volatilité, ce qui n'est pas étonnant étant donné le fonctionnement même de cette famille de produit. Aucune saisonnalité n'est évidente. Toute la difficulté sera donc de considérer les bonnes variables explicatives dans la modélisation.

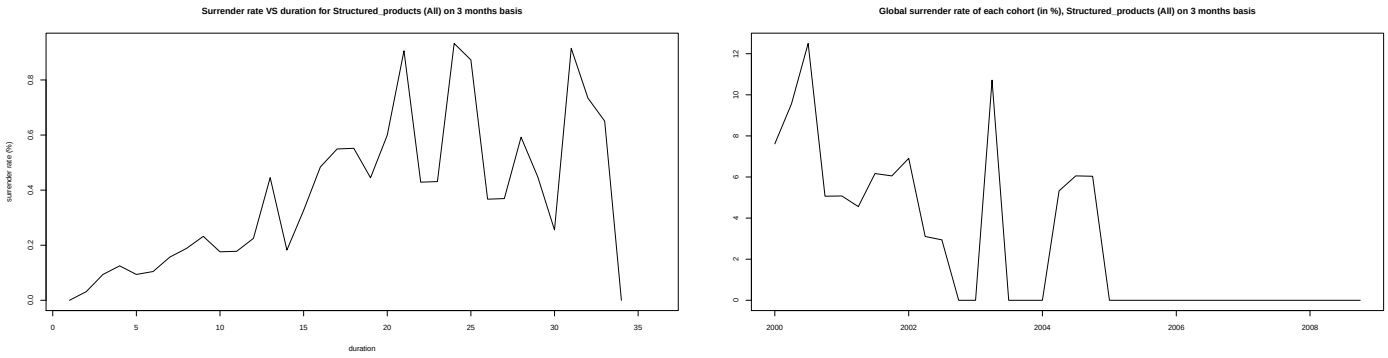


FIGURE 5.32 – A gauche : Taux de rachat en fonction de l'ancienneté de contrat. A droite : taux de rachat global par cohorte. Produits structurés.

Profil des rachats par ancienneté de contrat Le profil des rachats par ancienneté de contrat et le taux de rachat par cohorte (graphe 5.32), ainsi que le taux de rachat par date et par ancienneté de contrat (graphe 5.33) laisse présager une très forte hétérogénéité. Il est difficile de considérer que le taux de rachat est monotone en fonction de l'ancienneté, même si dans ce cas il semble qu'une tendance se dégage (plus d'assurés rachètent plus tard, ce qui peut paraitre surprenant d'ailleurs). En ce qui concerne le comportement des cohortes, il est très imprévisible (notons qu'il n'y a plus de souscription sur ce type de produit depuis début 2005). Nous ne sommes pas surpris de constater qu'il n'existe aucun profil type à dégager de ce type de produit, qui de par sa complexité et le fait qu'il dépende souvent uniquement des marchés rend les comportements aussi hétérogènes qu'imprévisibles. Notre objectif sera donc ici de ne pas trop compter sur des effets structurels qui à-priori n'aurait qu'un impact dérisoire sur la modélisation finale, et qui risque de "polluer" la modélisation. Ce postulat reste toutefois à vérifier dans l'application.

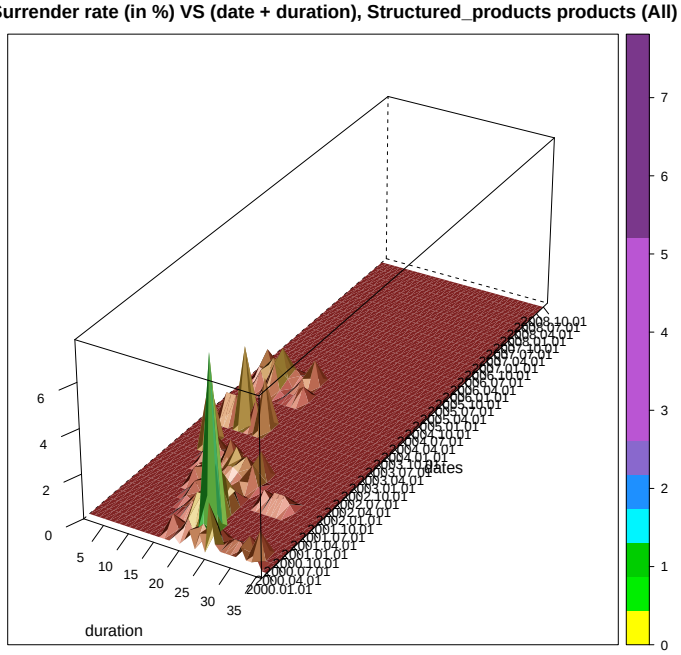


FIGURE 5.33 – Profil 3D du taux de rachat par date et par ancienneté de contrat (par trimestre), produit Structurés.

5.6.2 Sélection des variables : résultats par CART

Taux d'erreur de classification de l'arbre La classification des comportements de rachat des produits structurés est précise. Le taux d'erreur est seulement de 3,8 %, avec une spécificité de 99,8 % et une sensibilité de 40,3 %. Ces résultats peuvent quand même être trompeur lorsque l'on s'intéresse à ce type de produit en termes de modélisation. En effet, la dépendance aux marchés de cette famille de produit est telle que certaines variables qui apparaissent comme importantes dans ce classifieur peuvent finalement s'avérer inutiles à des fins de projection.

	Rachats non-observés	Rachat observés
Rachats non-prédits	8795	21
Rachats prédits	332	224

Importance des variables explicatives L'ancienneté du contrat, la richesse et l'âge sont les trois variables les plus discriminantes dans le processus de segmentation d'après la figure 5.34. Si nous considérons uniquement les variables catégorielles ou catégorisées, l'ancienneté et le réseau de distribution expliquent les décisions de rachat avec les meilleures prévisions. Nous allons voir que finalement aucune de ces variables explicatives n'est considérée dans le modèle prédictif des décisions de rachat, car leur introduction dégradaient clairement sa qualité.

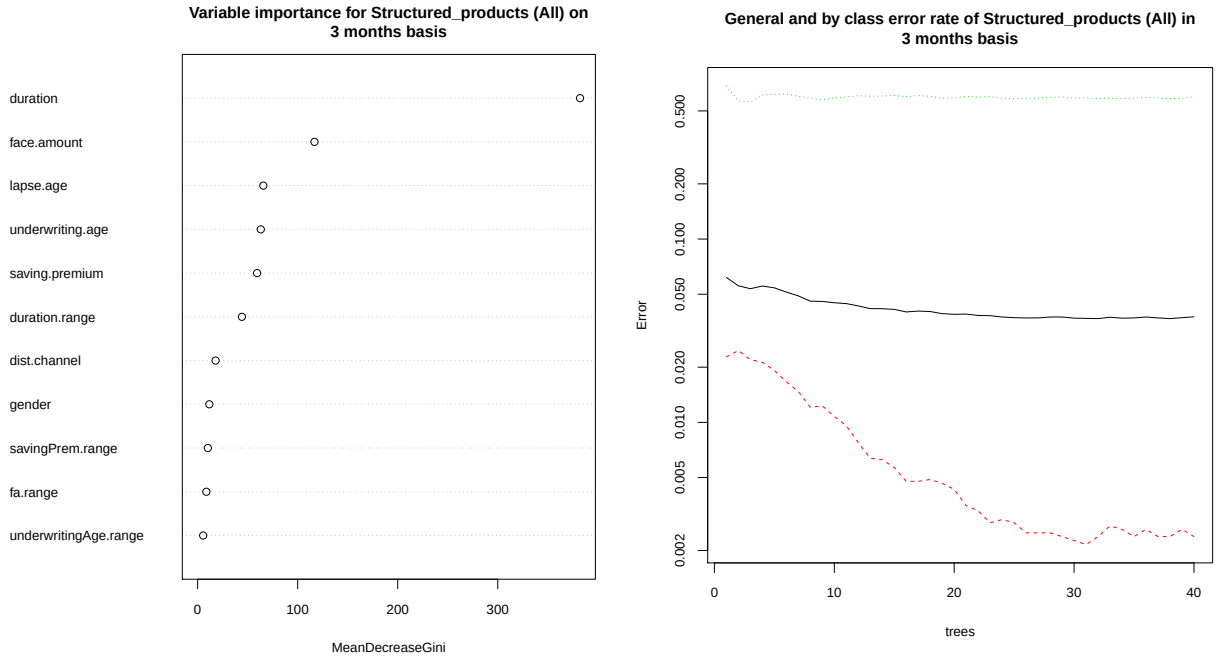
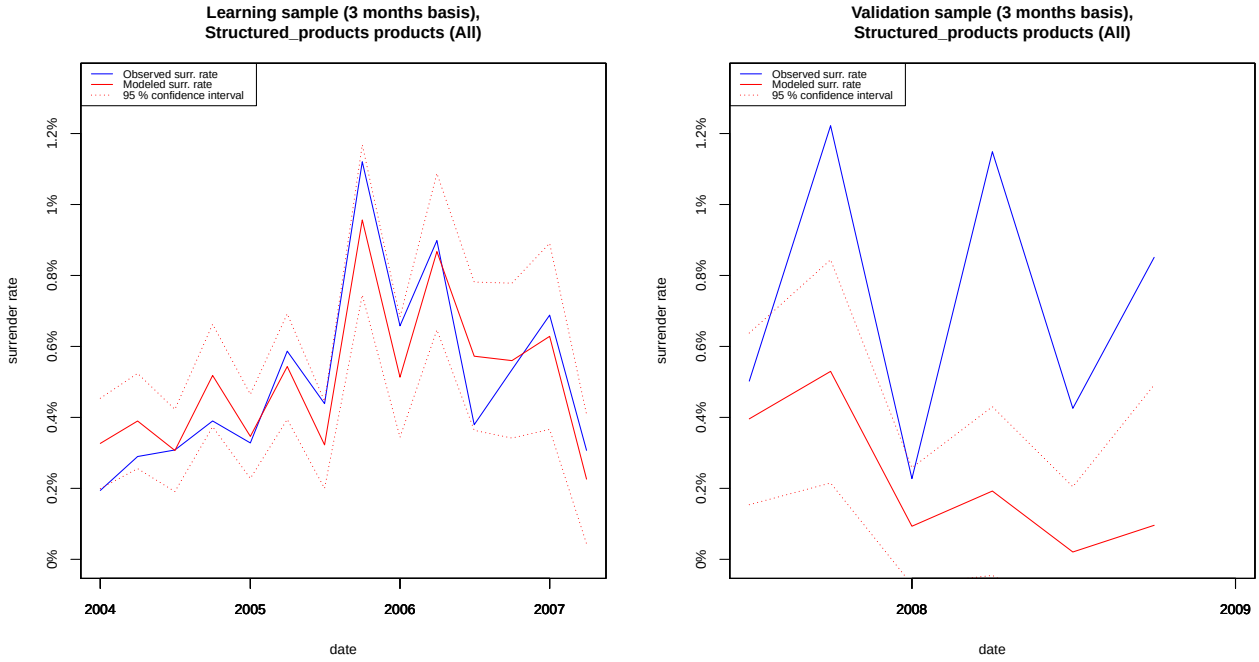


FIGURE 5.34 – Importance des variables explicatives, produit Structurés.

5.6.3 Modélisation et prévisions par mélange de GLM

Les prévisions par régression logistique dynamique du graphique 5.35 laissent à désirer. Autant le modèle performe bien sur la période d'apprentissage, autant il donne des résultats très mauvais sur la période de validation où presque toutes les observations sont en dehors de l'intervalle de confiance (lors du *back-testing*). Le changement de contexte économique qui impacte les comportements de rachat n'est donc pas bien capté par le modèle de régression logistique à une composante Avec un mélange de régressions logistiques, les prévisions sont nettement meilleures : quatre composantes suffisent à décrire précisément les comportements variables des assurés. Nous avons utilisé une méthode légèrement différente des approches considérées jusqu'ici, qui permet de mieux modéliser la corrélation entre individus. De par la nature du type de produit, la logique voudrait en effet que cette corrélation entre les comportements soit susceptible d'augmenter plus rapidement et plus intensément.

FIGURE 5.35 – Modélisation et prévision du taux de rachat des produits Structurés par régression logistique dynamique.



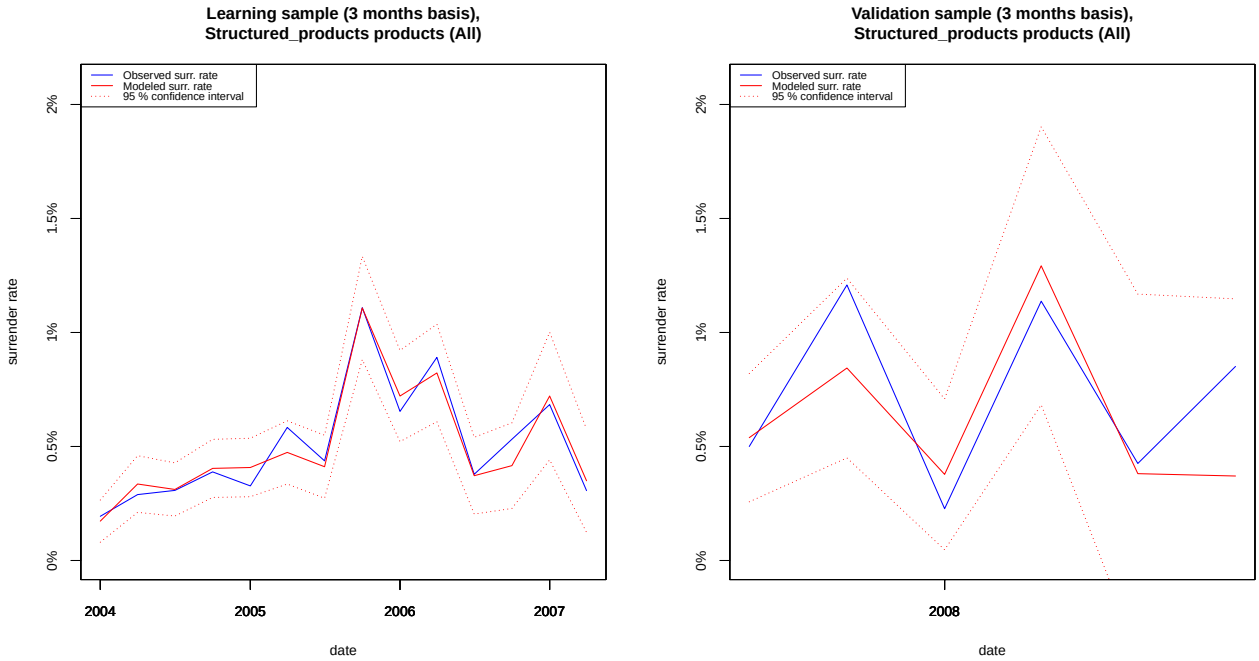


FIGURE 5.36 – Modélisation et prévision du taux de rachat par mélange de Logit, produits Structurés.

Impact des variables explicatives par les mélanges de Logit L'usage du modèle mélange est un peu spécifique pour les produits structurés. Nous n'avons introduit aucune variable explicative d'effet structurel (de type ancienneté de contrat, etc) car nous avons remarqué que les prévisions de comportement ne sont finalement pas du tout dictées par ces caractéristiques. Nos seuls facteurs de risque sont ceux liés à l'environnement économique et financier (taux 10Y et ibex 35), mais pris de manière spéciale comme le montrent les "boxplot" des effets de ces variables disponibles en annexe E.6.1 :

- effets *structurels* : aucun input.
- effets *conjoncturels* : introduits via l'Ibex 35 et le taux 10Y. Seul l'Ibex 35 joue sur la probabilité de rachat individuelle des assurés de chaque composante, avec un risque de base ("intercept") comparable entre composantes. Certains rachètent davantage lorsque l'indice croît, reflétant un comportement rationel (composantes 2, 3 et 4) ; alors que les individus appartenant à la composante 1 ont tendance à moins racheter dans un contexte haussier de l'indice.
- *corrélation* : l'originalité de l'approche pour les produits structurés consiste à pouvoir ajuster la taille (proportion) des composantes en fonction du contexte économique. Ceci est réalisé par l'introduction du taux 10 ans en variable explicative des poids des composantes (cf annexe E.13) : si le taux 10 ans augmente alors la probabilité d'appartenir à la composante 4 sera celle qui diminuera le plus, suivie de la composante 2 puis de la composante 1 (et inversement). En revanche plus d'individus adopteront le comportement représenté par la composante 3. Ainsi les assurés sont virtuellement autorisés à changer de composante chaque trimestre suivant l'économie.

Finalement, l'arbitrage du modèle dans la balance de la proportion de personnes adoptant une attitude plutôt rationnelle ou non permet d'obtenir un modèle dont les prévisions sont excellentes. L'aspect dynamique de la taille des composantes est ce qui nous a permis ici de trouver un résultat honorable sur des données relativement spécifiques.

5.7 Bilan

Le tableau 5.1 récapitule l'ensemble des résultats des modélisations ci-dessus. Cette représentation permet de faire ressortir des effets similaires sur des types de famille ressemblantes. La confiance que nous pouvons avoir en les estimations est résumée en colonne $Conf.$ (sur une échelle simplifiée allant de 1 à 3), et est basée sur la valeur et l'écart-type du coefficient calibré. Une confiance de 1 signifie que la calibration n'apparait pas robuste ; pour une estimation relativement satisfaisante une confiance de 2 est attribuée et enfin la valeur 3 représente une estimation excellente. La colonne $Nb.Comp.$ concerne le nombre de composantes retenu dans la modélisation, tandis que la qualité globale (toujours sur une échelle de 1 à 3) est évaluée suivant les graphiques de prévision en période de validation et les résultats des tests de Pearson et de Wilcoxon. Nous spécifions en dernière colonne quelques informations supplémentaires, en l'occurence la taille de l'échantillon d'apprentissage, la durée de la période de retour en arrière *delta* et la date de début pour la modélisation. Nous aurions évidemment pu jouer sur la valeur de ces différentes options à des fins d'optimisation des résultats (précision), mais notre but était de montrer que notre modélisation fonctionnait globalement sans avoir à adapter ces paramètres suivant les produits.

Nous tirons de ce tableau quelques conclusions intéressantes, notamment que nous

pourrions encore davantage regrouper (si tel était le besoin) les familles qui ont des modélisations proches :

- les contrats à **taux garanti** : grosso modo les "Ahorro", "Mixtos", "Pure Savings" et "Universal Savings" admettent le même type de modèle avec :
  - effets *structurels* (fixes entre composantes) dont toujours la saisonnalité et l'ancienneté du contrat, plus une (ou deux) variable(s) dépendante de la famille ;
  - effets *conjoncturels* : guidés par le taux long-terme ;
  - effets de *corrélation* : potentiels si les variables de richesse sont discriminantes.
- les contrats à rendement **non garanti** avec les "Index-Link" et "Unit-Link" :
  - effets *structurels* : pas ou peu de saisonnalité, l'ancienneté du contrat et une variable additionnelle dépendante de la famille ;
  - effets *conjoncturels* : plus intenses, guidés par les marchés financiers (Ibex 35) ;
  - effets de *corrélation* : potentiels si un scénario hyper stressé se réalise.
- les **produits structurés** : la complexité des produits peut expliquer ce comportement plus extrême :
  - effets *structurels* : pas d'effet clair donc inexistant dans la modélisation ;
  - effets *conjoncturels* : intense et dictés par le marché financier ;
  - effets de *corrélation* : introduits via le comportement du marché long-terme avec la taille de la composante risquée qui augmente si le marché se dégrade.

Famille	Covariables poids		Covariables composantes β fixés		Covariables composantes β variables		Nb. comp.	Qualité globale	Remarques (taille apprentissage, lookback period "delta")
	Nom	Conf.	Nom	Conf.	Nom	Conf.			
Ahorro	intercept	2	saisonnalité ancienneté fréquence prime âge souscription option PB	3 3 3 3 3	intercept taux 10Y	3 3	5	3	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2000
Index-Link	intercept	3	ancienneté âge souscription sexe	3 2 3	intercept ibex 35	3 2	2	3^+	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2000
Mixtos	intercept	1	saisonnalité ancienneté option PB	3 3 3	intercept prime risque taux 10Y	3 3 3	5	3^+	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2000
Pure Savings	intercept	2	saisonnalité ancienneté âge souscription	2 3 3	intercept ibex 35 taux 10Y	2 3 3	4	2^+	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2004
Unit-Link	intercept	2	saisonnalité ancienneté prime de risque	2 3 3	intercept ibex 35	3 3	5	3	apprentissage : 2/3 *delta* : 1 trimestre date début : 1/1/2000
Universal Savings	intercept	3	saisonnalité ancienneté âge souscription réseau distribution	3 3 2 3	intercept richesse	3 3	5	3^+	apprentissage : 2/3 *delta* : 1trimestre date début : 1/1/2004
Structured products	intercept Taux 10Y	1 2			intercept Ibex 35	3 3	4	3	apprentissage : 2/3 *delta* : 1, début : 1/1/2004

TABLE 5.1 – Tableau récapitulatif des modélisations retenues pour chaque famille de produits.

Conclusion du mémoire

Cette étude nous a permis de mieux comprendre comment appréhender les comportements de rachat grâce à une vision aggrégée, par définition plus complète et moins détaillée. L'étude empirique de différentes lignes de produits du portefeuille entier d'une entité d'AXA a été très instructive, et a servi de point de départ aux choix qui ont été adoptés par la suite. La volonté de comprendre, segmenter et modéliser les rachats à l'échelle de grandes familles de produit en utilisant des bases de données aggrégées induit des soucis de modélisation (à cause de l'hétérogénéité entre ces produits), mais se révèle clairement mieux adaptée pour une étude d'impact des rachats au niveau de la gestion actif-passif.

Le risque de rachat est un risque comportemental, donc par nature difficilement modélisable car dépendant de nombreux facteurs, aussi bien endogènes qu'exogènes. Suivant les positions de l'assureur et le type de produit, l'impact d'un scénario adverse des comportements de rachat peut être très conséquent. A ce titre, nous avons vu qu'il était primordial de prendre en compte la possible corrélation entre les comportements d'assurés lors de la modélisation dans une optique de gestion des risques (affinement du modèle interne). En effet cette dépendance entraîne une déformation de la distribution des rachats, qui provoque une hausse conséquente de la marge de risque. De plus, certaines caractéristiques clefs ne peuvent pas être négligées lors de la modélisation en cas d'évolution de la composition du portefeuille : nous pensons à l'ancienneté du contrat, à la richesse de l'assuré, au réseau de distribution. D'autres comme l'état de l'économie ou la réputation de l'entreprise, plus difficile à prendre en compte, jouent également un rôle prépondérant. Cette remarque suppose l'utilisation de modèles de régression, mais d'une manière que nous avons faite évoluer tout au long de ce projet. Comme une alternative aux précédentes approches vues en bibliographie, le coeur de ce travail s'articule autour d'une approche probabiliste invoquant une prise en compte spécifique des facteurs de risque structurels et conjoncturels. Nous évitons par cette approche l'hypothèse de rationalité et d'optimalité (au sens financier du terme) des assurés, et définissons une méthodologie **uniformisée** et semble-t-il **efficace** pour traiter le problème de la modélisation des rachats. Là est toute la nouveauté de notre étude, réplicable quelque soit le type de produit considéré par une adaptation logique des facteurs de risque correspondant à cette famille (dans les limites des familles de produit que nous étudions).

La principale "*leçon*" que nous retirons de ce projet d'étude est qu'il est inutile de prendre trop de facteurs de risque en compte : la saisonnalité, l'ancienneté de contrat, un troisième facteur de risque discriminant (endogène) et le contexte économique et financier suffisent en général à une bonne modélisation des comportements. La deuxième découverte est la manière de les considérer : les effets des facteurs idiosynchratiques doivent être fixés égaux entre les composantes des mélanges alors que le contexte

économique joue différemment sur les assurés. Cela permet à la fois de définir un cadre logique d'étude et de limiter le nombre de paramètres à estimer, donc la dimension de l'espace (permettant de meilleures prévisions). Les résultats de l'étude sont d'autant plus satisfaisants que la méthode de validation choisie (*back-testing*) fait intervenir de nouveaux contextes économiques en permanence, garantissant une bonne quantification des effets exogènes par la modélisation. Nous attirons l'attention du lecteur sur le fait que nous ne prétendons pas avoir trouvé la méthode idéale pour la modélisation des comportements de rachat ; simplement dans les différents cas testés, cette modélisation s'est avérée relativement fine et robuste. Un phénomène extrême qui ne serait pas dû à des mouvements sur les marchés financiers (exemple : politique de vente, image) pourrait avoir des conséquences dramatiques sur le taux de rachat mais ne serait évidemment pas prévu par notre modèle. Enfin, une approche que nous aurions aimé développer par la suite concerne les chaînes de Markov cachées qui nous paraissent bien adaptées à ce type de problème d'environnement changeant. La possibilité qu'elle offre de définir différents niveaux de risque de base ("regime switching") semble très intéressante de notre point de vue.

Bibliographie

Albert, F. S., Bragg, D. G. W. & Bragg, J. M. (1999), 'Mortality rates as a function of lapse rates', *Actuarial research clearing house* **1**. 16

Atkins, D. C. & Gallop, R. J. (2007), 'Re-thinking how family researchers model infrequent outcomes : A tutorial on count regression and zero-inflated models', *Journal of Family Psychology* . 15

Austin, P. C. (2007), 'A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting ami mortality', *Statistics in Medicine* **26**, 2937–2957. 31

Bacinello, A. R. (2005), 'Endogenous model of surrender conditions in equity-linked life insurance', *Insurance : Mathematics and Economics* **37**, 270–296. 10, 16

Bacinello, A. R., Biffis, E. & P., M. (2008), 'Pricing life insurance contracts with early exercise features', *Journal of Computational and Applied Mathematics* . 16

Balakrishnan, N. (1991), *Handbook of the Logistic Distribution*, Marcel Dekker, Inc. 31

Biard, R., Lefèvre, C. & Loisel, S. (2008), 'Impact of correlation crises in risk theory : Asymptotics of finite-time ruin probabilities for heavy-tailed claim amounts when some independence and stationarity assumptions are relaxed', *Insurance : Mathematics and Economics* **43**(3), 412 – 421. 48

Bluhm, W. F. (1982), 'Cumulative antiselection theory', *Transactions of Society of actuaries* **34**. 10

Box, G. & Cox, D. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society B* **26**, 211–252. 57

Breiman, L. (1994), Bagging predictors, Technical Report 421, Department of Statistics, University of California. 30

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* (24), 123–140. 30

Breiman, L. (1998), 'Arcing classifiers', *The Annals of Statistics* **26**(3), 801–849. 30

Breiman, L. (2001), 'Random forests', *Machine Learning* (45), 5–32. 30

Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees*, Chapman and Hall. 26, 28, 30, 31, 122

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. & Lindsay, B. (1994), 'The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family', *Annals of the Institute of Statistical Mathematics* **46**, 373–388. 62

Costabile, M., Massabo, I. & Russo, E. (2008), 'A binomial model for valuing equity-linked policies embedding surrender options', *Insurance : Mathematics and Economics* **40**, 873–886. 16

Cox, D. (1972), 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society : Series B* (34), 187–220. 17, 31

Cox, S. H. & Lin, Y. (2006), Annuity lapse rate modeling : tobit or not tobit ?, *in* 'Society of actuaries'. 15, 31

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, New Jersey : Princeton University Press. 63

Cummins, J. (1975), *An econometric model of the life insurance sector in the U.S. economy*, Lexington books, Health, Lexington/Mass u.a. 15

De Giovanni, D. (2007), Lapse rate modeling : A rational expectation approach, Finance Research Group Working Papers F-2007-03, University of Aarhus, Aarhus School of Business, Department of Business Studies. 16

Dempster, A., N.M., L. & D.B., R. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society* **39**, 1–38. 61

Denuit, M., Lefèvre, C. & Shaked, M. (1998), 'The s-convex orders among real random variables, with applications', *Math. Inequal. Appl.* **1**(4), 585–613. 128

Dutang, C. (2011), Regression models of price elasticity in non-life insurance, Master's thesis, ISFA. Mémoire confidentiel - AXA Group Risk Management. 15

Engle, R. & Granger, C. (1987), 'Cointegration and error-correction : Representation, estimation and testing', *Econometrica* (55), 251–276. 15

Fauvel, S. & Le Pévédic, M. (2007), Analyse des rachats d'un portefeuille vie individuelle : Approche théorique et application pratique, Master's thesis, ENSAE. Mémoire non confidentiel - AXA France. 15

Follmann, D. & Lambert, D. (1989), 'Generalizing logistic regression by non-parametric mixing', *Journal of the American Statistical Association* **84**, 295–300. 63

Fum, D., Del Missier, F. & A., S. (2007), 'The cognitive modeling of human behavior : Why a model is (sometimes) better than 10,000 words', *Cognitive Systems Research* **8**, 135–142. 9

Ghattas, B. (1999), 'Previsions par arbres de classification', *Mathematiques et Sciences Humaines* **146**, 31–49. 29, 121

Ghattas, B. (2000*a*), 'Aggregation d'arbres de classification', *Revue de statistique appliquee* **2**(48), 85–98. 30

Ghattas, B. (2000*b*), Importance des variables dans les methodes cart. GREQAM - Universite de Marseille. 30

Ghosh, J. & Sen, P. (1985), 'On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results', **2**, 789–806. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer. 63

Hilbe, J. M. (2009), *Logistic regression models*, Chapman and Hall. 26

Hin, H. K. & Huiyong, S. (2006), 'Structural prepayment risk behavior of the underlying mortgages for residential mortgage life insurance in a developing market', *Journal of Housing Economics* (15), 257–278. 17

Hosmer, D. W. & Lemeshow, S. (2000), *Applied Logistic Regression, 2nd ed.*, Wiley. 31

Kagraoka, Y. (2005), Modeling insurance surrenders by the negative binomial model. Working Paper 2005. 15, 31

Kim, C. (2005), 'Modeling surrender and lapse rates with economic variables', *North American Actuarial Journal* pp. 56–70. 15

Kim, C. N., Yang, K. H. & Kim, J. (2008), 'Human decision-making behavior and modeling effects', *Decision Support Systems* **45**, 517–527. 9

Kuen, S. T. (2005), 'Fair valuation of participating policies with surrender options and regime switching', *Insurance : Mathematics and Economics* **37**, 533–552. 10, 16

Lee, S., Son, Y.-J. & Jin, J. (2008), 'Decision field theory extensions for behavior modeling in dynamic environment using bayesian belief network', *Information Sciences* **178**, 2297–2314. 9

Lefèvre, C. & Utev, S. (1996), 'Comparing sums of exchangeable Bernoulli random variables', *J. Appl. Probab.* **33**(2), 285–310. 128

Lemmens, A. & Croux, C. (2006), 'Bagging and boosting classification trees to predict churn', *Journal of Marketing Research* **134**(1), 141–156. 34

Lindsay, B. & Lesperance, M. (1995), 'A review of semiparametric mixture models', *Journal of Statistical Planning and Inference* **47**, 29–99. 63

Lindstrom, M. & Bates, D. (1988), 'Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data', *Journal of the American Statistical Association* **83**, 1014–1022. 62

Liu, Y., Chawla, N., Harper, M., Shriberg, E. & Stolcke, A. (2006), 'A study in machine learning for unbalanced data for sentence boundary detection in speech.', *Computer Speech and Language* **20**(4), 468–494. 34

Loisel, S. (2008), 'From liquidity crisis to correlation crisis, and the need for quanls in enterprise risk management', pp. 75–77. in Risk Management : The Current Financial Crisis, Lessons Learned and Future Implications, Edited by the SOA, CAS and CIA. 48

Loisel, S. (2010), 'Contribution à la gestion quantitative des risques en assurance', *Habilitation Thesis, Université Lyon 1* . 56

Loisel, S. & Milhaud, X. (2011), 'From deterministic to stochastic surrender risk models : Impact of correlation crises on economic capital', *European Journal of Operational Research* . 9, 43, 56

Margolin, B., Kim, B. & Risko, K. (1989), 'The ames salmonella/microsome mutagenicity assay : issues of inference and validation', *Journal of the American Statistical Association* **84**, 651–661. 64

Marshall, A. & Olkin, I. (1979), *Inequalities : Theory of Majorization and Its Applications*, Academic Press, New York. 129

Martinussen, T. & Scheike, T. (2006), *Dynamic Regression Models for Survival Data*, Springer. 47

McCullagh, P. & Nelder, J. A. (1989), *Generalized linear models, 2nd ed.*, Chapman and Hall. 31

McLachlan, G. & Peel, D. (2000), *Finite Mixture Models*, Wiley Series In Porbability and Statistics. 57, 59, 62, 63

McNeil, A., Frey, R. & Embrechts, P. (2005), *Quantitative Risk Management*, Princeton Series In Finance. 9, 49

Milhaud, X., Gonon, M.-P. & Loisel, S. (2010), 'Les comportements de rachat en assurance vie en régime de croisière et en période de crise', *Risques* (83), 76–81. 9

Milhaud, X., Maume-Deschamps, V. & Loisel, S. (2011), 'Surrender triggers in life insurance : what main features affect the surrender behavior in a classical economic context ?', *Bulletin Francais d'Actuariat* **22**, ? 9

Nordahl, H. A. (2008), 'Valuation of life insurance surrender and exchange options', *Insurance : Mathematics and Economics* **42**, 909–919. 16

Outreville, J. F. (1990), 'Whole-life insurance lapse rates and the emergency fund hypothesis', *Insurance : Mathematics and Economics* **9**, 249–255. 15

Pan, X., Han, C. S., Dauber, K. & Law, K. H. (2006), 'Human and social behavior in computational modeling and analysis of egress', *Automation in Construction* **15**, 448–461. 9

Pearson, K. (1894), 'Contributions to the theory of mathematical evolution', *Philosophical Transactions of the Royal Society of London A* **185**, 71–110. 57

Pesando, J. (1974), 'The interest sensibility of the flow of funds through life insurance companies : An econometric analysis', *Journal Of Finance* **Sept**, 1105–1121. 15

Planchet, F. & Thérond, P. (2006), *Modèles de durée : applications actuarielles*, Economica (Paris). 47

Ramsay, J., Hooker, G. & Graves, S. (2009), *Functional Data Analysis with R and MATLAB*, Springer. 47

Ramsay, J. & Silverman, B. (2005), *Functional Data Analysis, Second Edition*, Spinger, Springer Series in Statistics. 47

Renshaw, A. E. & Haberman, S. (1986), 'Statistical analysis of life assurance lapses', *Journal of the Institute of Actuaries* **113**, 459–497. 15

Ruiz-Gazen, A. & Villa, N. (2007), 'Storms prediction : logistic regression vs random forest for unbalanced data', *Case Studies in Business, Industry and Government Statistics* **1**(2), 91–101. 34

Shen, W. & Xu, H. (2005), 'The valuation of unit-linked policies with or without surrender options', *Insurance : Mathematics and Economics* **36**, 79–92. 16

Stanton, R. (1995), 'Rational prepayment and the valuation of mortgage-backed securities', *Review of Financial* **8**, 677–708. 17

Teicher, H. (1963), 'Identifiability of finite mixtures', *Annals of Mathematical Statistics* **34**, 1265–1269. 64

Torsten, K. (2009), Valuation and hedging of participating life-insurance policies under management discretion, *in* 'Insurance : Mathematics and Economics Proceedings', Vol. 44, pp. 78–87. 17

Tsai, C., Kuo, W. & Chen, W.-K. (2002), 'Early surrender and the distribution of policy reserves', *Insurance : Mathematics and Economics* **31**, 429–445. 10, 16

Vandaele, N. & Vanmaele, M. (2008), 'Explicit portfolio for unit-linked life insurance contracts with surrender option', *Journal of Computational and Applied Mathematics* . 10, 16

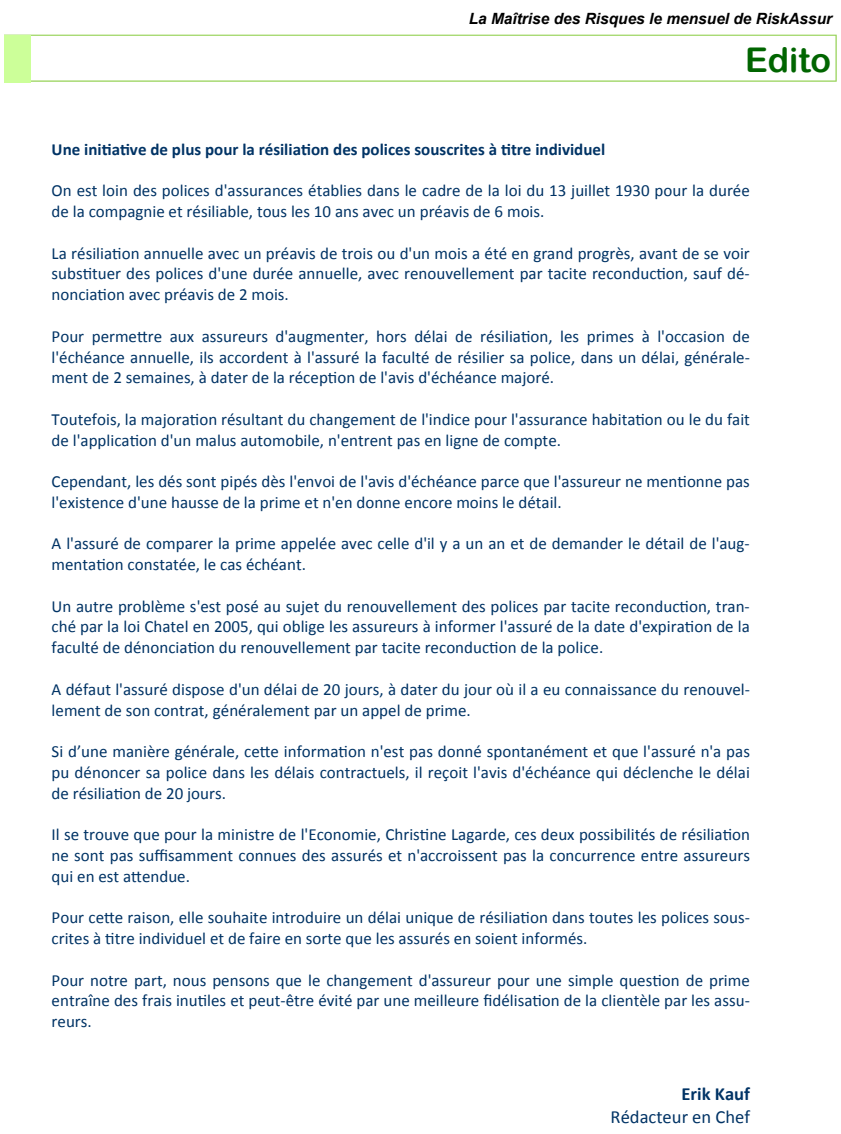
Viquerat, S. (2010), On the efficiency of recursive evaluations in relation to risk theory applications, PhD thesis. 52

Wang, P. (1994), Mixed Regression Models for Discrete Data, PhD thesis, University of British Columbia, Vancouver. 63

Wolfe, J. (1971), A monte carlo study of sampling distribution of the likelihood ratio for mixtures of multinormal distributions, Technical Report STB-72-2, San Diego : U.S. Naval Personnel and Training Research Laboratory. 63

Annexe A

Articles de presse



La Maîtrise des Risques le mensuel de RiskAssur

Edito

Une initiative de plus pour la résiliation des polices souscrites à titre individuel

On est loin des polices d'assurances établies dans le cadre de la loi du 13 juillet 1930 pour la durée de la compagnie et résiliable, tous les 10 ans avec un préavis de 6 mois.

La résiliation annuelle avec un préavis de trois ou d'un mois a été en grand progrès, avant de se voir substituer des polices d'une durée annuelle, avec renouvellement par tacite reconduction, sauf dénonciation avec préavis de 2 mois.

Pour permettre aux assureurs d'augmenter, hors délai de résiliation, les primes à l'occasion de l'échéance annuelle, ils accordent à l'assuré la faculté de résilier sa police, dans un délai, généralement de 2 semaines, à dater de la réception de l'avis d'échéance majoré.

Toutefois, la majoration résultant du changement de l'indice pour l'assurance habitation ou le du fait de l'application d'un malus automobile, n'entrent pas en ligne de compte.

Cependant, les dés sont pipés dès l'envoi de l'avis d'échéance parce que l'assureur ne mentionne pas l'existence d'une hausse de la prime et n'en donne encore moins le détail.

A l'assuré de comparer la prime appelée avec celle d'il y a un an et de demander le détail de l'augmentation constatée, le cas échéant.

Un autre problème s'est posé au sujet du renouvellement des polices par tacite reconduction, tranché par la loi Chatel en 2005, qui oblige les assureurs à informer l'assuré de la date d'expiration de la faculté de dénonciation du renouvellement par tacite reconduction de la police.

A défaut l'assuré dispose d'un délai de 20 jours, à dater du jour où il a eu connaissance du renouvellement de son contrat, généralement par un appel de prime.

Si d'une manière générale, cette information n'est pas donné spontanément et que l'assuré n'a pas pu dénoncer sa police dans les délais contractuels, il reçoit l'avis d'échéance qui déclenche le délai de résiliation de 20 jours.

Il se trouve que pour la ministre de l'Economie, Christine Lagarde, ces deux possibilités de résiliation ne sont pas suffisamment connues des assurés et n'accroissent pas la concurrence entre assureurs qui en est attendue.

Pour cette raison, elle souhaite introduire un délai unique de résiliation dans toutes les polices souscrites à titre individuel et de faire en sorte que les assurés en soient informés.

Pour notre part, nous pensons que le changement d'assureur pour une simple question de prime entraîne des frais inutiles et peut-être évité par une meilleure fidélisation de la clientèle par les assureurs.

Erik Kauf
Rédacteur en Chef

FIGURE A.1 – RiskAssur du 29/04/2011, changement législatif potentiel pour la résiliation des polices.

GESTION D'ACTIFS

EN BREF

SÉBASTIEN BARBE REJOINT FEDERAL FINANCE. Sébastien Barbe passe de Rothschild & Cie Gestion à Federal Finance. Il sera détaché auprès de Schelcher Prince Gestion en qualité de directeur général délégué. Cette nomination s'inscrit dans le projet de rapprochement entre Federal Finance et Schelcher Prince Gestion.

ÉPARGNE SALARIALE : INFORMATION, FORMATION ET CONSEIL, MAÎTRES MOTS. L'Autorité des marchés financiers (AMF) a publié son rapport sur l'épargne salariale et l'actionnariat salarié. Le groupe de travail présidé par Jacques Delmas-Marsalet a rendu ses conclusions. *L'article complet sur lesechos.fr*

H_2O AM : BIENTÔT 1 MILLIARD D'EUROS. Les fondateurs de H_2O Asset Management, adossé à Natixis AM, Bruno Crastes et Vincent Chailley, n'avaient pas imaginé qu'ils auraient plus de 500 millions d'euros d'encours huit mois après leur lancement, et autant en promesses d'ici à trois mois. *Lire « Les Echos » du 11 février*

LUNDI 14 FÉVRIER 2011

RISQUE

Plus d'une société de gestion a trébuché dans son histoire. Les derniers mois l'ont encore illustré, avec les affaires Gartmore ou AXA Rosenberg, dont la réputation a été entachée. Les clients pardonnent rarement, surtout quand le gérant n'a pas daigné jouer la transparence.

Quand la réputation d'une société de gestion est mise en cause

Dès que la réputation est engagée, la pérennité de l'entreprise est en danger. C'est particulièrement vrai dans le monde de la gestion d'actifs, où les histoires d'« accident industriel » ou d'affaires fâcheuses ne manquent pas.

Le britannique Gartmore a ainsi dû renoncer à son indépendance après la suspension d'un de ses gérants vedettes par le régulateur et le départ de nombreux autres. Quelques mois après son introduction en Bourse, il a été racheté par son compatriote Henderson, en janvier dernier. Les gérants mis en cause dans les affaires de « market timing » et de « late trading » en 2003 ont mis du temps à se remettre, après avoir été sanctionnés. Le plus durement touché, Putnam (alors filiale de Marsh & McLennan) a été repris fin 2006 par Power Corporation. Janus Capital a dû lui aussi prendre les mesures qui s'imposaient (lire ci-dessous).

Une préoccupation essentielle

Selon un sondage réalisé par Ernst &

pesante. Il existe en effet plus d'un piège susceptible de faire perdre la face à une société de gestion. La crise financière l'a bien illustré. Quand des gérants ont fermé brutalement leurs sicav monétaires dynamiques peu après le déclenchement de la crise des « subprimes », les dédommagements furent longs à venir, et la gestion monétaire en a souffert. Quand Lehman Brothers a

« Les investisseurs ne veulent plus de boîte noire. »

VINCENT PUCHE, PRÉSIDENT DU CABINET DE CONSULTANTS INSTI7

fait faillite, en septembre 2008, il a fragilisé des établissements bancaires et, par ricochet, leurs filiales de gestion d'actifs. « *Les institutions ont pris en compte dans les critères d'évaluation la manière dont les groupes ont été affectés durant la crise finan-*

POURQUOI LA GESTION DES RISQUES EST-ELLE IMPORTANTE DANS UNE SOCIÉTÉ DE GESTION ?

RANG 2010		RANG 2009
1	SURCROÎT DE RÉGLEMENTATION	2
2	ÉVITER LE RISQUE DE RÉPUTATION	1
3	OPTIMISER LES FONDS PROPRES ET LES LIQUIDITÉS	6
4	ACCROISSEMENT DES RESPONSABILITÉS VIS-À-VIS DES TIERS	4
5	PRESSION DES CLIENTS	3
6	MAINTENIR LES MEILLEURES PRATIQUES DE MARCHÉS	5

censées choisir les meilleurs gérants après avoir procédé à des « due diligences ».

L'affaire AXA Rosenberg n'a pas été non plus sans dommage collatéral sur la société elle-même (lire l'interview ci-dessous) ni sur l'ensemble de la gestion quantitative. « *Elle a montré que les modèles ne sont pas des pierres philosophales, qu'il était bon qu'ils soient transparents. Les investisseurs ne veulent plus de boîte noire* », relève Vincent Puche, président du cabinet de consultants Insti7. « *Le "feed-back" des investisseurs a été tout d'abord désastreux pour les stratégies actions quantitatives en général. Beaucoup sont sortis*, a noté Olivier Jesequel. *Nous posons plus de questions pour savoir quels sont les risques qu'une situation similaire se produise chez d'autres gérants, quelles actions ont été menées de leur part pour se prémunir d'erreurs de programmation...* »

En sortie de crise, le risque de réputation est très prégnant dans les

fait l'objet de procédures judiciaires », fait remarquer Vincent Puche, qui a noté que les clients revenaient rarement vers un gérant avec lesquels ils ont connu des déboires.

Attention au risque humain

Chez les gérants, la vigilance est de tous les instants, surtout pour les équipes de conformité ou de contrôle des risques. Tout est visé, de la qualité des communications sur les produits à la passation des ordres en passant par les contreparties (règlement-livraison et compensation) et les prestataires externes. Le risque humain nécessite une attention toute particulière. « *Nous devons détecter le plus en amont les risques liés à chaque acteur, que ce soit dans les interventions sur les marchés ou les déclarations à la presse qui pourraient être mal interprétées et qui engagent la société* », détaille Frédéric Beugin, responsable de la conformité et du contrôle interne chez Robeco Gestions. La gestion des

Figure A.2 – Les Echos du 14 février 2011 - Classement de l'importance des types de risques pour une société de gestion. Le risque de réputation est classé en 2ème position en 2010 (cf tableau), après avoir été classé 1er en 2009.

Annexe B

Annexes sur les méthodes de segmentation

B.1 Méthode CART

B.1.1 Etapes de construction de l'arbre

Les différentes étapes de construction de l'arbre sont résumées dans le schéma suivant :

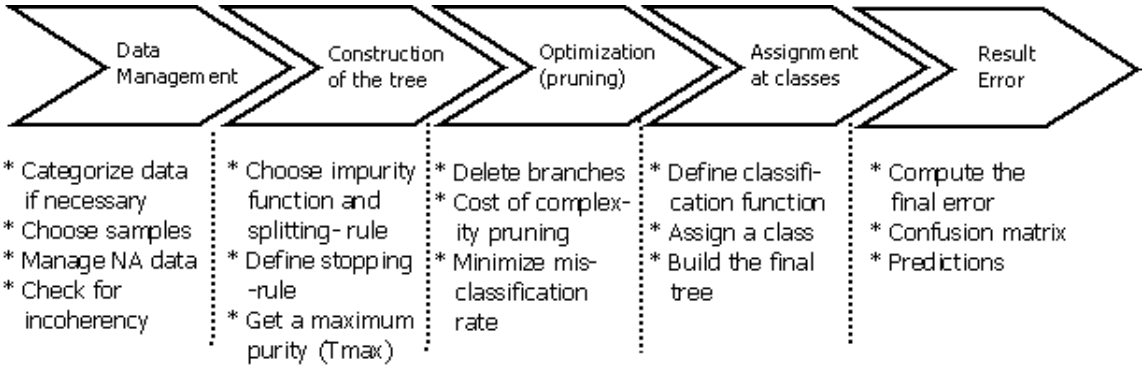


FIGURE B.1 – Etapes chronologiques de la procédure CART

Nous détaillons également par un dessin (en figure B.2) la division d'un noeud, donnant lieu à de nouvelles branches et un gain d'homogénéité.

B.1.2 Choix du paramètre de complexité

`rpart()` élague l'arbre par K validations croisées (K=10 par défaut) sur chaque arbre élagué (nous avons pris K=10). Les assurés sont choisis aléatoirement dans le processus de validations croisées, c'est pourquoi la *cptable* peut différer légèrement entre deux simulations. Sur la table B.1, *relerror* mesure l'erreur d'apprentissage et donne la qualité d'ajustement de l'arbre, *xerror* mesure le taux de mauvaise classification des 10 validations croisées et est considérée comme un meilleur estimateur de l'erreur réelle. *xstd* est l'écart type de *xerror*. L'arbre optimal minimise $err = xerror + xstd$. Si deux arbres ont la même erreur *err*, nous choisissons le plus petit. La table B.1 permet de

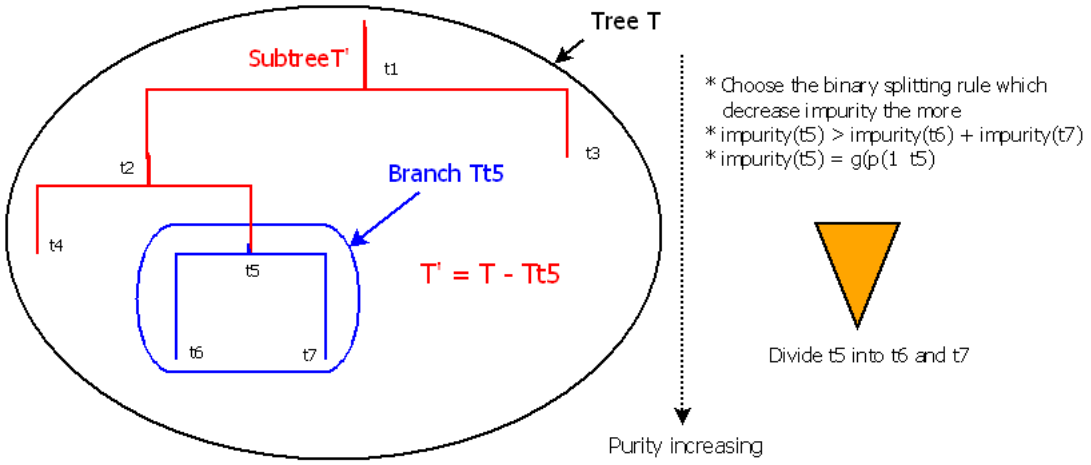


FIGURE B.2 – Construction d'un arbre binaire

tracer l'erreur d'apprentissage en fonction du paramètre de complexité et de la taille de l'arbre (voir figure B.3).

Remark 1 *Quelques commentaires sur la lecture de ce tableau :*

- *le troisième arbre avec deux divisions correspond à* $\alpha \in]2.30, 3.10]$,
- `R` *normalise l'erreur, ce qui explique que l'erreur de la racine soit de 100% (1). La vraie erreur de la racine peut être obtenue en affichant l'arbre (ici elle est de 45.465%),*
- *l'arbre maximal* T_{max} *(non élagué) retourné par défaut par la fonction* `rpart()` *correspond à la dernière ligne de la cptable.*

CP	nsplit	rel error	xerror	xstd
3.3981e-01	0	1.000	1.000	0.0084
3.0539e-01	1	0.660	0.660	0.0077
5.9982e-03	2	0.354	0.361	0.0062
7.8237e-04	5	0.336	0.337	0.0061
5.2158e-04	10	0.331	0.333	0.0060
4.5638e-04	15	0.328	0.333	0.0060
3.9119e-04	19	0.326	0.333	0.0060
3.6945e-04	21	0.325	0.333	0.0060
3.2599e-04	32	0.319	0.333	0.0060
3.1295e-04	34	0.318	0.333	0.0060
2.6079e-04	39	0.317	0.332	0.0060
2.1733e-04	53	0.31360	0.334	0.0060

CP	nsplit	rel error	xerror	xstd
1.9559e-04	59	0.312	0.332	0.0060
1.8255e-04	68	0.310	0.332	0.0060
1.3040e-04	73	0.309	0.332	0.0060
1.0432e-04	82	0.308	0.332	0.0060
9.7796e-05	88	0.307	0.333	0.0060
8.6930e-05	97	0.306	0.334	0.0060
6.5198e-05	100	0.306	0.334	0.0060
4.3465e-05	117	0.305	0.337	0.0061
3.7256e-05	132	0.304	0.339	0.0061
3.2599e-05	139	0.304	0.340	0.0061
2.6079e-05	159	0.303	0.340	0.0061
0.0000e+00	174	0.303	0.341	0.0061

TABLE B.1 – Paramètres de complexité, *cptable*

B.1.3 Plus loin dans la théorie des CART

Spécification des règles binaires

Criterion 1 *Ces règles dépendent seulement d'un seuil μ et d' une variable x_l, $1 \leq l \leq d$:*

- *$x_l \leq \mu, \mu \in \mathbb{R}$ dans le cas d'une variable continue ordonnée (si nous avons m valeurs distinctes pour x_l, l'ensemble des valeurs possibles card(D) vaut M - 1) ;*
- *$x_l \in \mu$ où μ est un sous-ensemble de $\{\mu_1, \mu_2, ..., \mu_M\}$ et les μ_m sont les modalités de la variable catégorielle (dans ce cas le cardinal du sous-ensemble D des règles binaires vaut $2^{M-1} - 1$).*

Qu'est ce qu'une fonction d'impureté ?

Definition 1 *Une fonction d'impureté est une fonction réelle g définie sur un ensemble de probabilités discrètes d'un ensemble fini :*

$$g : (p_1, p_2, ..., p_J) \rightarrow g(p_1, p_2, ..., p_J),$$

symétrique en p_1, p_2, ..., p_J et qui vérifie :

1. *le maximum de g est à l'équiprobabilité : $argmax\ g(p_1, p_2, ..., p_J) = \left(\frac{1}{J}, \frac{1}{J}, ..., \frac{1}{J}\right)$,*
2. *le minimum de g est obtenu par les "dirac" : $argmin\ g(p_1, p_2, ..., p_J) \in \{e_1, ..., e_J\}$, où e_j est le j^{eme} élément dans la base canonique de$\mathbb{R}^J$.*

Différentes fonctions d'impureté

D'habitude nous considérons les fonctions suivantes (qui satisfont le critère de concavité) :

- impur(t) = - $\sum_{j=1}^{J} p(j|t)\ \ln(p(j|t))$;
- impur(t) = $\sum_{j \neq k} p(j|t)\ p(k|t)$ (index de Gini)

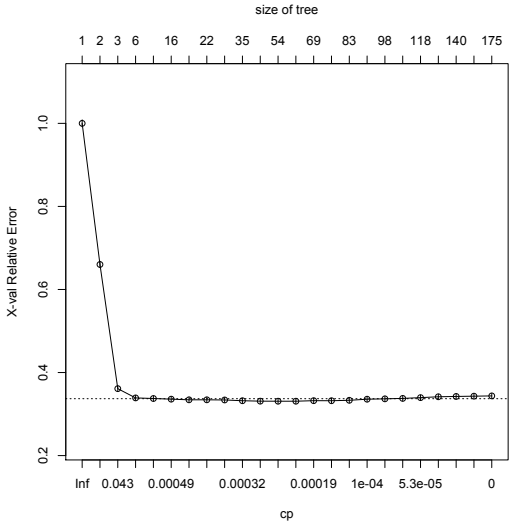


FIGURE B.3 – L'estimateur du taux de mauvaise classification de l'arbre optimal en fonction du paramètre de complexité cp (ou α). T_{max} contient ici 175 feuilles et correspond à $cp = 0$. Remarquez la forme avec un forte pente négative suivie d'un plateau, puis une légère remontée de l'erreur.

Remark 2 *Dans une approche variance,*

- *l'index de Gini est aussi égal à* $1 - \sum_j p_j^2$ *;*
- *nous utilisons également la twoing rule : choisir* Δ *qui maximise* $\frac{p_L p_R}{4}\Big[\sum_j |p(j|t_L) - p(j|t_R)|\Big]^2$ *;*
- *dans un problème avec une réponse binaire, l'index de Gini se réduit à* $impur(t) = 2p(1|t)p(2|t)$.

Commentaires sur l'erreur de prévision

Nous pouvons écrire de manière formelle l'expression de la portion d'observations mal classées par la fonction *class* suivant l'estimation choisie de l'erreur de prévision :

– l'estimate "resubstitution" :

$$\hat{\tau}(class) = \frac{1}{N} \sum_{(x_n, j_n) \in \epsilon} \mathbb{1}\{class(x_n, \epsilon) \neq j_n\} \tag{B.1}$$

– l'estimation par échantillon de validation : quasiment comme dans $(B.1)$:

$$\hat{\tau}^{ts}(class) = \frac{1}{N'} \sum_{(x_n, j_n) \in W} \mathbb{1}\{class(x_n, \epsilon) \neq j_n\} \tag{B.2}$$

– l'estimation par validations croisées :

$$\hat{\tau}^{cv}(class) = \frac{1}{N} \sum_{k=1}^{K} \sum_{(x_n, j_n) \in \epsilon_k} \mathbb{1}\{class(x_n, \epsilon^k) \neq j_n\} \tag{B.3}$$

Remarquons aussi que

$$\begin{aligned}
\mathbb{E}[\hat{\tau}(class)] &= \mathbb{E}\left[\tfrac{1}{N} \sum_{(x_n, j_n) \in \epsilon} \mathbb{1}\{class(x_n, \epsilon) \neq j_n\}\right] \\
&= \frac{1}{N} \sum_{(x_n, j_n) \in \epsilon} \mathbb{E}[\mathbb{1}\{class(x_n, \epsilon) \neq j_n\}] \\
&= P(class(X, \epsilon) \neq Y) = \tau(class).
\end{aligned}$$

et que tous les estimateurs présentés ci-dessus sont non-biaisés :
$\mathbb{E}[\hat{\tau}(class)] = \mathbb{E}[\hat{\tau}^{cv}(class)] = \mathbb{E}[\hat{\tau}^{ts}(class)]$
L'erreur de prévision et le taux de mauvaise classification sont deux concepts différents. L'erreur de mauvaise classification est l'erreur dans les noeuds de l'arbre alors que l'erreur de prévision est liée à la classification finale de la variable d'intérêt et est calculée une fois l'arbre construit.

Par défaut, `R` calcule un estimateur par validations croisées de l'erreur d'apprentissage. Ce sont les résultats du tableau des paramètres de complexité. Toutefois cette procédure de validations croisées ne correspond pas à la fameuse technique de validations croisées dans la théorie du reéchantillonage. La première calcule l'arbre optimal pour une taille donnée en minimisant l'erreur d'apprentissage alors que la dernière permet d'obtenir une estimation plus réaliste de l'erreur de prévision mais ne traite pas le problème qui est de trouver un arbre optimal.

Pénalisation de la mauvaise classification

Les méthodes en structure d'arbre ont subi beaucoup de critiques à cause de la taille des arbres finaux sélectionnés en pratique et de l'usage de l'estimation par resubstitution (cf 2.1.1). Le coût de mal classer une observation n'est souvent pas le même pour toutes les classes dans les applications, d'où l'idée de pénaliser la mauvaise classification d'une observation (par rapport à sa classe observée, apprentissage supervisé) par un facteur positif.

Definition 2 *Le coût de mauvais classement d'une observation est défini par*

$$\Gamma : C \times C \to \mathbb{R}_+ \text{ , such that}$$
$$\Gamma(i|j) \geq 0 \text{ and } \Gamma(i|i) = 0$$

Définissons ainsi

- la probabilité de mal classer une observation par $P_{class}(i|j) = P(class(x, \epsilon) = i \mid j)$ (la fonction *class* classe x dans la classe i au lieu de la classe j),
- $\tau_{class}(j) = \sum_i \Gamma(i|j) P_{class}(i|j)$: le coût moyen de mauvaise classification.

Nous obtenons $\tau_{class} = \tau(T)$ et

$$\tau(T) = \sum_j \pi(j)\tau_{class}(j) = \frac{1}{N} \sum_j N_j \tau_{class}(j)$$

Ghattas (1999) définit dans ce contexte la fonction de classification pénalisée d'assignation d'une classe à un noeud terminal t :

$$class(x, \epsilon) = \underset{i \in C}{argmin} \sum_{j \in C} \Gamma(i|j)\ p(j|t) \tag{B.4}$$

D'après $(B.4)$, l'estimation du taux de mauvaise classification est maintenant

$$r(t) = \min_{i \in C} \sum_{j \in C} \Gamma(i|j)\ p(j|t)$$

Sachant que $\tau(t) = r(t)p(t)$, le taux de mauvaise classification par substitution de l'arbre T est donné par

$$\hat{\tau}(T) = \sum_{t \in \tilde{T}} \hat{\tau}(t). \tag{B.5}$$

Corollary 1 *L'estimateur $\hat{\tau}(T)$ du taux de mauvaise classification de l'arbre s'abaissent à chaque division, et ce quelque soit la division. Ainsi, si nous notons T_s l'arbre obtenu par division de T à une feuille, nous avons*

$$\hat{\tau}(T_s) \leq \hat{\tau}(T) \tag{B.6}$$

Soient t_L et t_R les descendants du noeud t dans l'arbre T_s.
D'après $(B.5)$ et $(B.6)$,

$$\begin{aligned} \sum_{t \in \tilde{T}_s} \hat{\tau}(t) &\leq \sum_{t \in \tilde{T}} \hat{\tau}(t) \\ \sum_{t \in \tilde{T}} \hat{\tau}(t) - \hat{\tau}(t) + \hat{\tau}(t_L) + \hat{\tau}(t_R) &\leq \sum_{t \in \tilde{T}} \hat{\tau}(t) \\ \hat{\tau}(t_L) + \hat{\tau}(t_R) &\leq \hat{\tau}(t) \end{aligned} \tag{B.7}$$

Elagage de l'arbre

Le problème d'un arbre final trop complexe qui "overfit" les données peut être résolu assez facilement. La solution consiste à appliquer ces deux idées plutôt que d'essayer de trouver la bonne règle d'arrêt des divisions (qui n'est pas la bonne approche) :

1. ne pas arrêter la construction de l'arbre et obtenir le plus grand arbre T_{max} ; puis l'élaguer par étape jusqu'à la racine (le critère d'élagage et de recombinaison de l'arbre est beaucoup plus important que le critère de division) ;
2. utiliser de meilleurs estimateurs du vrai taux de mauvaise classification pour sélectionner l'arbre de bonne taille parmi les sous-arbres élagués (utiliser les validations croisées ou l'échantillon témoin/test pour cela).

L'idée est de chercher des sous-arbres de T_{max} avec un taux de mauvaise classification minimum. Elaguer une branche T^t d'un arbre T signifie supprimer tous les descendants du noeud t dans T. L'arbre élagué résultantest noté $T' = T - T^t$, et $T' < T$.
D'après ($B.7$) nous avons

$$\hat{\tau}(t) \geq \hat{\tau}(T^t). \tag{B.8}$$

T_{max} contient tellement de noeuds qu'un nombre incalculable de manières d'élaguer l'arbre jusqu'à la racine existe, ce qui nous amène à définir un critère pour la sélection de la procédure d'élagage qui donne le "meilleur" sous-arbre. Le critère naturel de comparaison pour les arbres de même taille est l'erreur de mauvaise classification : l'algorithme d'élagage commence par T_{max} et élague progressivement jusqu'à obtenir la racine de telle manière qu'à chaque étape de l'élagage le taux de mauvaise classification soit aussi faible que possible. Ce travail fournit une suite d'arbres de plus en plus petits : $T_{max} > T_1 > T_2 > ... > T_{root}$. ($T_{root}$ est le noeud racine, sans aucune division).
D'après ($B.6$), remarquons que : $T_1 < T_{max} \Rightarrow \hat{\tau}(T_{max}) \leq \hat{\tau}(T_1)$. L'erreur de l'arbre maximal est toujours inférieure ou égale à l'erreur de l'arbre élagué, le but étant de diminuer le nombre de feuilles de T_{max}. Une idée naturelle consiste à pénaliser un grand nombre de feuilles dans l'arbre final, par l'introduction dans l'erreur d'un terme de coût de complexité. Le nouveau taux de mauvaise classification ou *cost-complexity measure* devient :

$$\hat{\tau}_\alpha(T) = \hat{\tau}(T) + \underbrace{\alpha\ Card(\tilde{T})}_{\text{complexity term}}, \text{ where } \alpha > 0, \tag{B.9}$$

où $Card(\tilde{T})$ est le nombre de noeuds terminaux de T.
En fait nous désirons juste trouver le sous-arbre $T(\alpha) \leq T_{max}$ qui minimise $\tau_\alpha(T)$ pour chaque α :

$$\tau_\alpha(T(\alpha)) = \min_{T \leq T_{max}} \tau_\alpha(T) \tag{B.10}$$

Pour les questions d'existence et d'unicité de l'arbre $T(\alpha)$, voir l'ouvrage de Breiman et al. (1984).
α est clairement lié à la taille de l'arbre final élagué ; si α est petit alors la pénalité associée à un grand nombre de feuilles est petite et l'arbre $T(\alpha)$ sera grand. Les cas extrêmes sont :

- $\alpha = 0$: chaque feuille contient une seule observation (T_{max} très grand). Toutes les observations sont bien classées et $\tau(T_{max}) = 0$. T_{max} minimise $\tau_0(T)$;
- $\alpha \to +\infty$: la pénalité pour le nombre de feuilles est grande et le sous-arbre qui minimise l'erreur sera la racine !

Algorithm 1 *Pour connaître les branches à élaguer et le α optimal associé,*

1. *Soient les feuilles t_L et t_R les descendants immédiats du noeud parent t ; en commençant par T_{max}, nous cherchons la division qui n'a pas donné de diminution de l'erreur, i.e. pour laquelle $\hat{\tau}(t) = \hat{\tau}(t_L) + \hat{\tau}(t_R)$ (voir (B.7)). Elaguer t_L et t_R, et recommencer de même jusqu'à ce que ce ne soit plus possible. Nous obtenons $T_1 < T$;*

2. *Pour T_1^t branche de T_1, définissons $\hat{\tau}(T_1^t) = \sum_{t \in \tilde{T}_1^t} \hat{\tau}(t)$. D'après (B.8), les noeuds non-terminaux t de l'arbre T_1 satisfont la propriété : $\hat{\tau}(t) > \hat{\tau}(T_1^t)$ (pas d'égalité grâce à la première étape).*

3. *Notons $\{t\}$ la sous-branche de T_1^t qui consiste en l'unique noeud $\{t\}$, card($\{t\}$) = 1.*
   Ainsi, $\hat{\tau_\alpha}(\{t\}) = \hat{\tau}(t) + \alpha$ et

$$\hat{\tau_\alpha}(T_1^t) = \hat{\tau}(T_1^t) + \alpha \; Card(\tilde{T}_1^t) \tag{B.11}$$

   Nous avons vu que $\hat{\tau}(T_1^t) < \hat{\tau}(\{t\})$, mais l'introduction d'un terme de complexité fait que cette inégalité avec $\hat{\tau_\alpha}$ n'est pas toujours respectée. Tant que $\hat{\tau_\alpha}(T_1^t) < \hat{\tau_\alpha}(\{t\})$ il est inutile d'élaguer, mais il existe un seuil α_c tel que $\hat{\tau_{\alpha_c}}(T_1^t) = \hat{\tau_{\alpha_c}}(\{t\})$. On a donc

$$\begin{aligned} \hat{\tau}(T_1^t) + \alpha_c \; Card(\tilde{T}_1^t) &= \hat{\tau}(t) + \alpha_c \\ \alpha_c &= \frac{\hat{\tau}(t) - \hat{\tau}(T_1^t)}{Card(\tilde{T}_1^t) - 1} \end{aligned}$$

   Tant que $\alpha < \alpha_c$, il n'est pas nécessaire d'élaguer l'arbre au noeud t, mais dès que $\alpha = \alpha_c$ l'élagage de cette sous-branche est intéressante car l'erreur est équivalente et l'arbre est plus simple ;

4. *Faire ceci pour tous les noeuds t de T_1 et choisir le noeud t dans T_1 qui minimise la quantité α_c. Soit $\alpha_1 = \alpha_c$. En élaguant T_1 au noeud t, nous obtenons $T_2 = T_1 - T_1^t$. Répéter 3. et 4. récursivement avec T_2, obtenez α_2 et ainsi de suite jusqu'à la racine.*

Au final, nous obtenons par construction (avec les cas extrêmes) une suite $\alpha_1 < \alpha_2 < ... < \alpha_{root}$ qui correspondaux arbres élagués $T_1 > T_2 > ... > T_{root}$. T_{root} est juste le noeud racine.
Pour définir l'arbre optimal de cette suite, ($B.10$) nous dit que le meilleur arbre élagué est celui avec le taux de mauvaise classification minimum.

B.2 La régression logistique

B.2.1 Résultats numériques de l'analyse statique

Les coefficients de régression, leur écart-type, la confiance que nous pouvons avoir dans l'estimation de ces coefficients et leur effet sont disponibles dans la table B.2. Les coefficients de régression de l'analyse dynamique du début du chapitre... ne sont pas donnés ici car ils n'ont pas vraiment d'intérêt (l'analyse logistique dynamique avait pour but de montrer que les prévisions n'étaient pas robustes).

B.2.2 Un peu de théorie

La modélisation "logit" est pertinente car nous voulons étudier un événement binaire (le rachat), or la régression logistique analyse des données issues de loi binomiale de la forme $Y_i \sim B(n_i, p_i)$, avec n_i le nombre d'expériences de bernoulli et p_i la probabilité de succès (rachat ici). Si nous notons Y la variable à expliquer (i.e. la décision de rachat), nous avons

$$Y = \begin{cases} 1, & \text{si l'assuré rachète sa police,} \\ 0, & \text{sinon.} \end{cases}$$

Nous pouvons maintenant adapter l'équation de régression logistique à notre contexte et nous obtenons la probabilité de rachat p :

$$\begin{aligned} logit &= \ln\left(\frac{P[Y=1|X_0=x_0,...,X_k=x_k]}{P[Y=0|X_0=x_0,...,X_k=x_k]}\right) \\ &= \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \end{aligned}$$

Finalement,

$$\left.\begin{aligned} \Phi(logit(p)) &= \Phi(\Phi^{-1}(p)) = p \\ \Phi(logit(p)) &= \Phi(\beta_0 + \textstyle\sum_{j=1}^{k} \beta_j X_j) \end{aligned}\right\} (1)$$

$(1) \Rightarrow p = \Phi(\beta_0 + \sum_{j=1}^{k} \beta_j X_j)$.
Cette écriture permet de comprendre plus facilement l'expression de la fonction de vraisemblance en 2.2.2.

B.2.3 L'algorithme de Newton-Raphson

Maxmiser la fonction de log-vraisemblance (**??**) amène à la résolution du système (k + 1) équations

$$\begin{cases} \dfrac{\partial l}{\partial \hat{\beta}_0} &= \sum_{i=1}^{n} Y_i - \Phi(\beta_0 + \sum_{j=1}^{k} \beta_k X_k) = 0 \\ \dfrac{\partial l}{\partial \hat{\beta}_j} &= \sum_{i=1}^{n} X_{ij}(Y_i - \Phi(\beta_0 + \sum_{j=1}^{k} \beta_k X_k)) = 0 \end{cases}$$

$\forall j = 1, ..., k$.
Le problème est que les solutions n'admettent pas de formules fermées et l'utilisation d'un algorithme d'optimisation est alors indispensable. Souvent l'algorithme de Newton-Raphson (basé en fait sur un développement de Taylor à l'ordre 1) est utilisé à cette fin. En SAS et en R, cet algorithme est inclus et lance le processus itératif suivant :

$$\beta^{(i+1)} = \beta^{(i)} - \left(\frac{\partial^2 \ln(L(\beta))}{\partial\beta\partial\beta'}\right)^{-1} \times \left(\frac{\partial \ln(L(\beta))}{\partial\beta}\right) \tag{B.12}$$

Lorsque la différence entre $\beta^{(i+1)}$ et $\beta^{(i)}$ est plus petite qu'un certain seuil (disons par exemple 10^{-4}), les itérations s'arrêtent et nous obtenons la solution finale.

B.2.4 Estimation de la matrice de covariance

La matrice de variance Z des coefficients $\hat{\beta}$ s'écrit

$$\begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & \cdots & Cov(\hat{\beta}_0, \hat{\beta}_k) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_1) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_k, \hat{\beta}_0) & Cov(\hat{\beta}_k, \hat{\beta}_1) & \cdots & Var(\hat{\beta}_k) \end{pmatrix} \tag{B.13}$$

et est estimatée par l'inverse de la matrice d'information de Fisher, qui vaut

$$I(\beta) = -\mathbb{E}\Big[\frac{\partial^2 \ln(L(\beta))}{\partial\beta\partial\beta'}\Big].$$

Un des caractéristiques intéressantes est que le dernier terme de cette équation est déjà calculé par l'algorithme de Newton-Raphson, ce qui permet d'estimer les coefficients de régression et leur matrice de covariance simultanément.
Comme d'habitude, l'estimateur par maximum de vraisemblance $\hat{\beta}$ converge asymptotiquement vers une loi normale de moyenne la vraie valeur de β et de variance l'inverse de la matrice de Fisher $I(\beta)$.
Le terme dans l'espérance est appelé la *Hessienne* et est également utilisé dans les tests de significativité des coefficients de régression β.

B.2.5 Statistique de déviance et tests

Evaluation statistique de la régression

Pour vérifier la pertinence du modèle, nous utilisons la statistique du test du ratio de vraisemblance : l'hypothèse nulle de ce test est $\beta_1 = \beta_2 = ... = \beta_k = 0$ (H_0) ;
Et l'hypothèse alternative est "au moins un des coefficients de régression n'est pas nul" (H_1).
Soit l(β) la log-vraisemblance du modèle de régression logistique avec $k+1$ coefficients de régression, et l(β_0) la log-vraisemblance du modèle de régression logistique le plus simple (avec seulement l'ordonnée à l'origine β_0), la statistique du ratio de vraisemblance vaut

$$\Lambda = 2 \times \Big(l(\beta) - l(\beta_0)\Big). \tag{B.14}$$

Cette statistique suit une loi du χ^2_k à k degrès de liberté (d.f.).
Si la p-valeur est plus petite que le niveau de confiance que nous nous accordons, alors le modèle est globalement significatif et H_0 est rejetée.
Plus intuitivement, les statisticiens utilisent parfois le coefficient R^2 (ou coefficient de MC Fadden) : $R^2 = 1 - \dfrac{l(\beta)}{l(\beta_0)}$.
Un coefficient R^2 proche de 0 signifie que le ratio de vraisemblance est proche de 1, et donc que la log-vraisemblance du modèle complet est proche de celle du modèle le plus simple. Ainsi il n'est pas très utile d'introduire des variables explicatives supplémentaires pour la modélisation. A l'opposé, si R^2 est proche de 1, alors il y a une grande différence en termes de vraisemblance entre les deux modèles et il est intéressant de considérer le modèle complet qui est bien meilleur.

Pertinence et significativité d'une variable explicative

L'idée de ce test est de comparer la valeur du coefficient estimé β_j (associé à la variable explicative X_j) à sa variance, elle-même extraite de la matrice hessienne.
L'hypothèse nulle (H_0) de ce test est : $\beta_j = 0$; et l'hypothèse alternative (H_1) est donnée par : $\beta_j \neq 0$.
Nous utilisons la statistique de Wald qui suit une distribution du χ^2_1 pour réaliser ce test : $\Lambda = \dfrac{\hat{\beta_j}^2}{Var(\hat{\beta_j})}$.
Choisissons par exemple un seuil de confiance de 5%, et notons $\chi^2_{95\%}(1)$ le 95^{eme} percentile de la loi du chi-deux à 1 d.f. H_0 est vraie si le ratio est plus petit que ce quantile, sinon nous rejetons H_0.

Coef. (var. type)	modality : correspondance	coefficient estimate	std error	p-value	effect
β_0 (continuous)		10.63398	1.48281	7.42e-13	> 0
$\beta_{duration}$ (categorical)	1 : [0,12] (in month)	0 (reference)			nul
	2 :]12,18]	-1.31804	0.15450	$< 2e-16$	< 0
	3 :]18,24]	-2.66856	0.14016	$< 2e-16$	< 0
	4 :]24,30]	-2.75744	0.14799	$< 2e-16$	< 0
	5 :]30,36]	-3.09368	0.14294	$< 2e-16$	< 0
	6 :]36,42]	-3.54961	0.15080	$< 2e-16$	< 0
	7 :]42,48]	-3.72161	0.14980	$< 2e-16$	< 0
	8 :]48,54]	-4.10431	0.15772	$< 2e-16$	< 0
	9 : > 54	-5.49307	0.14037	$< 2e-16$	< 0
$\beta_{premium\ frequency}$ (categorical) (in month)	Monthly	0 (reference)			nul
	Bi-monthly	0.92656	0.62071	0.135504	> 0
	Quarterly	-0.03284	0.10270	0.749148	< 0
	Half-yearly	-0.22055	0.16681	0.186128	< 0
	Annual	0.43613	0.10690	4.51e-05	> 0
	Single	-0.28494	0.38155	0.455177	< 0
$\beta_{underwriting\ age}$ (categorical)	1 : [0,20[(years old)	0 (reference)			nul
	2 : [20,30[	0.28378	0.13912	0.041376	> 0
	3 : [30,40[	-0.01146	0.13663	0.933163	< 0
	4 : [40,50[	-0.26266	0.14077	0.062054	< 0
	5 : [50,60[	-0.42098	0.15136	0.005416	< 0
	6 : [60,70[	-0.66396	0.19531	0.000675	< 0
	7 : > 70	-0.75323	0.23417	0.001297	< 0
$\beta_{face\ amount}$ (categorical)	1 :	0 (reference)			nul
	2 :	-5.79014	1.46592	7.82e-05	< 0
	3 :	-7.14918	1.46631	1.08e-06	< 0
$\beta_{risk\ premium}$ (categorical)	1 :	0 (reference)			nul
	2 :	0.36060	0.11719	0.002091	> 0
	3 :	0.26300	0.14041	0.061068	> 0
$\beta_{saving\ premium}$ (categorical)	1 :	0 (reference)			nul
	2 :	0.93642	0.13099	8.74e-13	> 0
	3 :	1.32983	0.14955	$< 2e-16$	> 0
$\beta_{contract\ type}$ (categorical)	PP con PB	0 (reference)			nul
	PP sin PB	-16.79213	114.05786	0.882955	< 0
	PU con PB	-7.48389	1.51757	8.16e-07	< 0
	PU sin PB	-12.43284	1.08499	$< 2e-16$	< 0
β_{gender}	Female	0 (reference)			nul
	Male	-0.08543	0.04854	0.078401	< 0

TABLE B.2 – Estimations des coefficients de la régression logistique pour les contrats Mixtes.

Annexe C

Annexes sur les crises de corrélation

C.1 Les ordres stochastiques pour une comparaison qualitative

Soit $M_{(p,p_0)}$ le nombre d'assurés qui rachètent leur contrat dans le modèle où $P(J_1 = 1) = p_0$ et $P(I_1 = 1) = p$. Examinons comment les valeurs de p et p_0 affectent la distribution du taux de rachat conditionnel. Nous utilisons pour cela l'ordre stochastique $s-$convex (Lefèvre & Utev (1996) et Denuit et al. (1998)). Sachant deux variables aléatoires Y et Z, pour $s = 1, 2, \ldots$, nous avons

$$X \leq^{\mathcal{D}}_{s-cx} Y \text{ if } E[\phi(Y)] \leq E[\phi(Z)] \text{ pour tout } s\text{-fonction convexe } \phi : \mathcal{D} \to \mathbb{R}, \quad \text{(C.1)}$$

donc grosso modo pour n'importe quelle fonction ϕ sur $\mathcal{D}$ dont la s-eme dérivée existe et satisfait $\phi^{(s)} \geq 0$. Les $s-1$ premiers moments de Y et Z sont d'ailleurs nécessairement égaux. L'ordre $\leq^{\mathcal{D}}_{1-cx}$ correspond à l'ordre stochastique individuel $\leq_1$, $\leq^{\mathcal{D}}_{2-cx}$ est l'ordre convexe usuel $\leq_2$ (qui implique en particulier que $Var(Y) \leq Var(Z)$). De plus, X est dit plus petit que Y dans l'ordre convexe croissant (noté $\leq_{icx}$) si

$$E\left(f(X)\right) \leq E\left(f(Y)\right)$$

pour toute fonction convexe croissante telle que l'espérance existe.

Proposition 1 *Lorsque le paramètre de corrélation est fixé, le nombre de rachats est stochastiquement croissant en p : pour $p_0 \in (0,1)$ fixé, si $p < p'$ alors*

$$M_{(p,p_0)} \leq_1 M_{(p',p_0)}.$$

Preuve : Cette proposition découle de résultats élémentaires sur les ordres stochastiques impliquant des distributions Binomiales et de Bernouilli. □

Proposition 2 *Lorsque la probabilité individuelle de rachat p est fixé, le paramètre de corrélation induit un ordre 2-convex du nombre de rachats : pour $p \in (0,1)$ fixé, si $p_0 < p_0'$ alors*

$$M_{(p,p_0)} \leq_2 M_{(p,p_0')}.$$

Preuve : Sachant que le nombre de comportements moutonniers N vaut k, le nombre total de rachats est

$$M_{(p,k)} = k.I_0 + 0.I_1^{\perp} + 0.I_2^{\perp} + ... + 0.I_k^{\perp} + 1.I_{k+1}^{\perp} + ... + 1.I_n^{\perp}.$$

Sachant que $N = k^{'}$ avec $k \leq k^{'}$, nous avons

$$M_{(p,k')} = k^{'}.I_0 + 0.I_1^{\perp} + ... + 0.I_k^{\perp} + ... + 0.I_{k'}^{\perp} + 1.I_{k'+1}^{\perp} + ... + 1.I_n^{\perp}.$$

Nous pouvons comparer les deux variables aléatoires $M_{(p,k)}$ et $M_{(p,k')}$ par un ordre de majorisation (voir par exemple Marshall & Olkin (1979)). Notons $Z^{\downarrow} = (z_1^{\downarrow}, \ldots, z_K^{\downarrow})$ le vecteur des mêmes composantes que Z (quelconque) classées en ordre décroissant. Connaissant deux vecteurs $Y = (y_1, \ldots, y_K)$ et $Z = (z_1, \ldots, z_K)$ de taille $K \geq 1$ tels que

$$\sum_{i=1}^{K} y_i = \sum_{i=1}^{K} z_i,$$

rappelons que Z est dit majorant Y si pour tout $j \leq K$,

$$\sum_{i=1}^{j} y_i^{\downarrow} = \sum_{i=1}^{j} z_i^{\downarrow}.$$

D'après Marshall & Olkin (1979), si le vecteur $\alpha = (\alpha_0, ..., \alpha_n)$ est plus petit que le vecteur $\beta = (\beta_0, ..., \beta_n)$ dans l'ordre de majorisation partielle, et si les X_i sont i.i.d., alors nous obtenons l'ordre convexe suivant :

$$\sum_i \alpha_i X_i \leq_2 \sum_i \beta_i X_i.$$

Nous posons $X_i = I_i^{\perp}$ et

$$(\alpha_0, ..., \alpha_n) = (k, \underbrace{0, ..., 0}_{k\ fois}, 1, ..., 1) \text{ et } (\beta_0, ..., \beta_n) = (k^{'}, \underbrace{0, ..., 0}_{k'\ fois}, 1, ..., 1).$$

Pour $k \leq k'$, le vecteur $(\beta_0, ..., \beta_n)$ majore clairement le vecteur $(\alpha_0, ..., \alpha_n)$. De plus, la variable aléatoire $(N \sim Bin(n, p_0))$ is stochastiquement croissante en p_0. Nous pouvons donc conclure que pour $p_0 \leq p_0'$,

$$M_{(p,p_0)} \leq_2 M_{(p,p_0')},$$

où $M_{(p,p_0)}$ est le nombre d'assurés qui rachètent quand la probabilité de se comporter en mouton vaut p_0 et quand la probabilité individuelle de rachat est p. □

Proposition 3 *Dans le modèle où p et p_0 augmentent simultanément Δr, Δr induit un ordre convexe croissant du nombre de rachats : soit M (resp. M') le nombre de rachats quand $\Delta r = x$ (resp. $\Delta r = x'$). Si $x < x'$ alors nous avons*

$$M \leq_{icx} M'.$$

Preuve : Nous utilisons les mêmes arguments que dans les preuves des propositions 1 et 2. En combinant ces arguments, le résultat est immédiat car quand Δr croît, p et p_0 augmentent. □

Ces propositions permettent l'interprétation de deux résultats pratiques en termes de gestion de risque :

- La Proposition 1 implique que l'espérance, la *VaR* à n'importe quel seuil $\alpha \in (0, 1)$ et les primes stop-loss $E[(M - m)_+]$ pour $0 \leq m \leq n$ sont croissantes en p et en Δr.

– La Proposition 2 dit que la variance et les primes stop-loss $E\left[(M-m)_+\right]$ pour $0 \leq m \leq n$ sont croissantes en p_0 (à p fixé). Elle montre aussi que si $p_0 < p_0'$, il existe un niveau $\alpha_0 \in (0,1)$ tel que pour $\alpha > \alpha_0$,

$$VaR_\alpha\left(M_{(p,p_0)}\right) < VaR_\alpha\left(M_{(p,p_0')}\right)$$

car le critère de Karlin-Novikov énonce dans ce cas que les fonctions de répartition de $M_{(p,p_0)}$ et $M_{(p,p_0')}$ ne peuvent se croiser qu'une fois.

Il n'est évidemment pas surprenant d'apprendre qu'augmenter le paramètre de corrélation et/ou la probabilité marginale de rachat provoque une plus grande mesure de risque en général, mais le but ici est aussi de déterminer l'importance de l'impact de cette corrélation sur le besoin en capital économique et sur la valeur du compte (notion introduite par la suite) dans un contexte réel.

Annexe D

Annexes des tests de la modélisation mélange

D.1 Résultats des tests de validation de modèle prédictif

D.1.1 Test de Pearson

```
> normality.test(obj, "Pearson")

Pearson chi-square normality test

data:  validation.residuals
P = 0.8, p-value = 0.8495
```

D.1.2 Test de Mann-Whitney-Wilcoxon

```
> distribution.test(obj, "Wilcoxon-Mann-Whitney")

Wilcoxon rank sum test

data:  obj[[2]] and obj[[3]]
W = 50, p-value = 1
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -0.04310852  0.03889232
sample estimates:
difference in location
         -0.0001792725
```

Annexe E

Annexes sur les applications

E.1 Famille de produits Ahorro

E.1.1 Données formatées pour la modélisation

```
"issue.date";"termination.date";"line.of.business";"contract.type";"PB.guarantee";"product.no";
"1";"1999-01-01";"2008-01-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"462";
"2";"1999-01-01";"2008-01-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"313";
"3";"1999-01-01";"2008-01-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"328";
"4";"1999-01-01";"2008-01-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"462";
"5";"1999-01-01";"2004-11-01";"Saving";"Ahorro_PP_con_PB";"with_PB";"328";

"premium.frequency";"gender";"lapse.age";"underwriting.age";"underwritingAge.range";"face.amount";
"unique";"Male";NA;49;"2";9562.04;
"highly.periodic";"Female";NA;32;"1";44560.44;
"highly.periodic";"Male";NA;49;"2";23064.41;
"highly.periodic";"Female";NA;41;"2";36986.28;
"highly.periodic";"Male";26;20;"1";11706.51;

"fa.range";"risk.premium";"riskPrem.range";"saving.premium";"savingPrem.range";"duration";
"high.face.amount";0;"low.risk.premium";601.01;"middle.saving.premium";35.7802197802198;
"high.face.amount";0;"low.risk.premium";490.16;"middle.saving.premium";35.7802197802198;
"high.face.amount";0;"low.risk.premium";1667.22;"high.saving.premium";35.7802197802198;
"high.face.amount";0;"low.risk.premium";1050.42;"high.saving.premium";35.7802197802198;
"high.face.amount";0;"low.risk.premium";336.7;"low.saving.premium";23.4175824175824;

"duration.range";"lapse.reason";"lapse.bit";"surrender.bit"
"high.duration";"In force";"0";"0"
"high.duration";"In force";"0";"0"
"high.duration";"In force";"0";"0"
"high.duration";"In force";"0";"0"
"high.duration";"Surrender";"1";"1"
```

E.1.2 Taux de rachat global par cohort et boxplot des coefficients

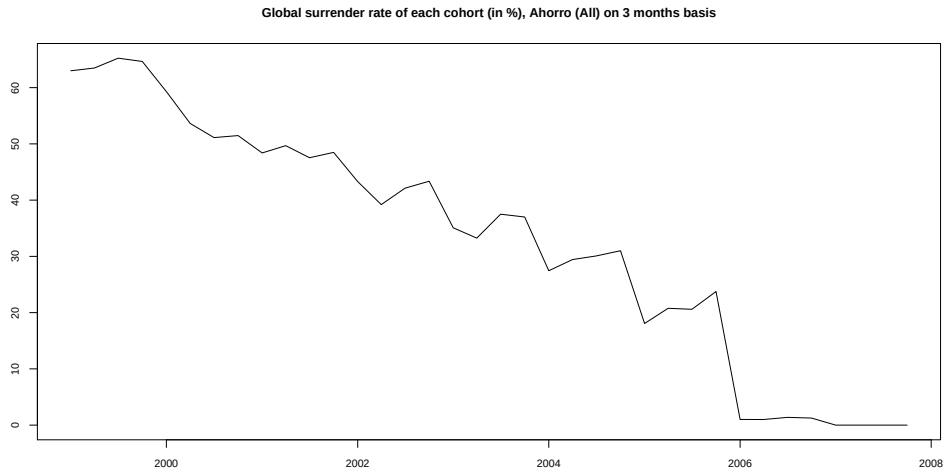


FIGURE E.1 – Pourcentage global de rachat par cohorte pour les produits Ahorro.

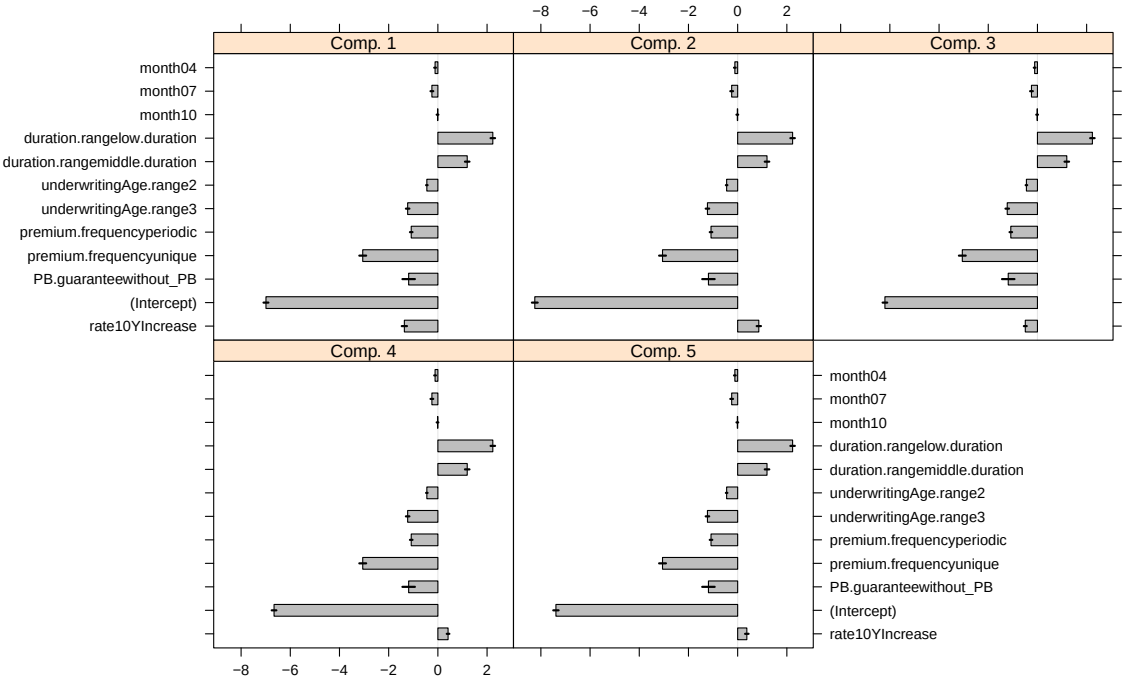


FIGURE E.2 – Coefficients de régression des composantes, produits Ahorro.

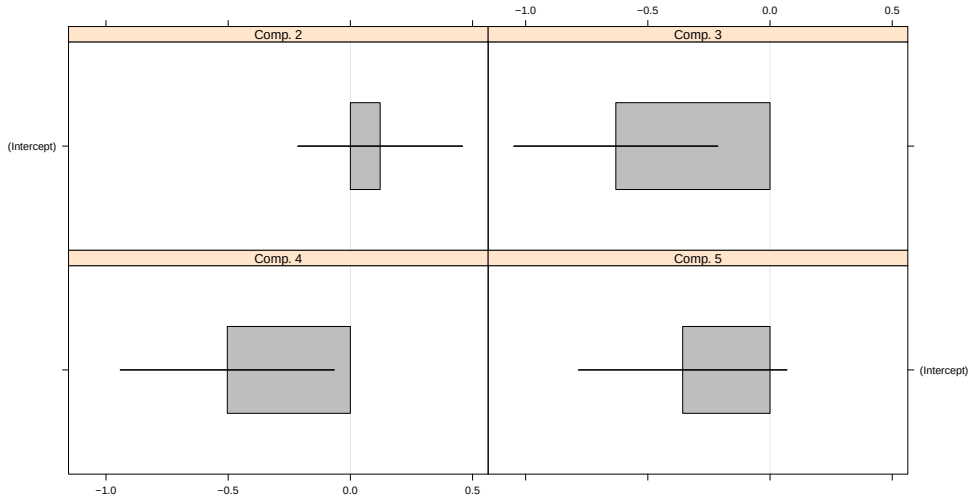


FIGURE E.3 – Coefficients de régression estimés des poids des composantes, produits Ahorro.

E.2 Famille de produits Unit-Link

E.2.1 Boxplot des coefficients

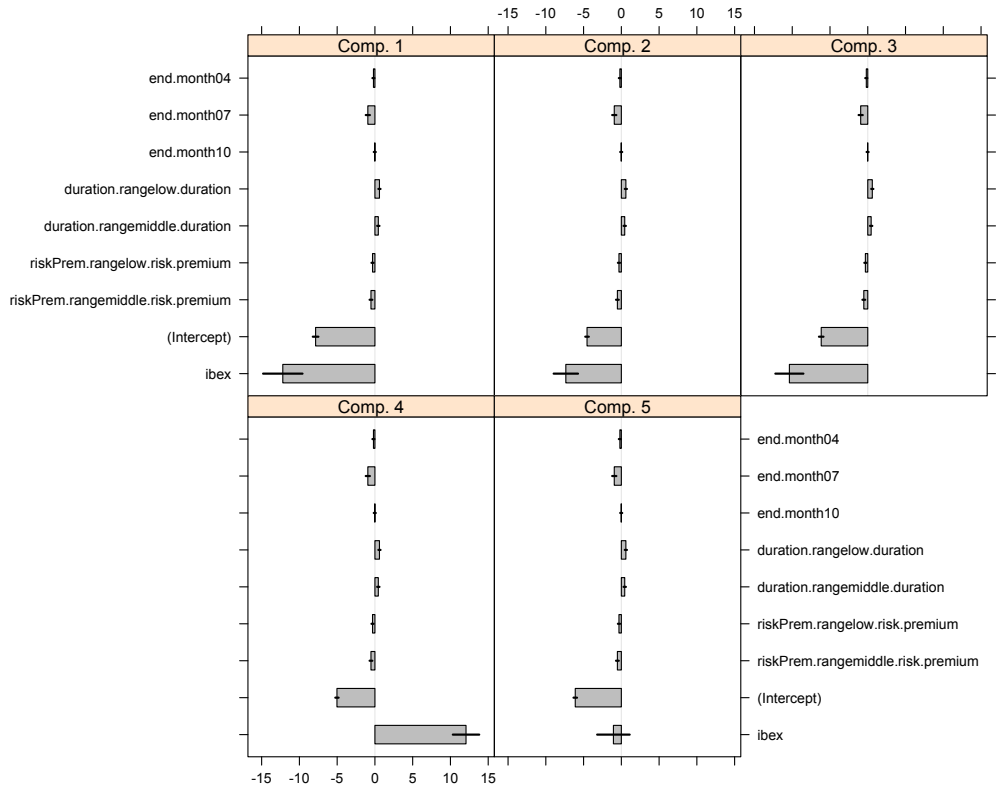


FIGURE E.4 – Coefficients de régression des composantes, produits UC.

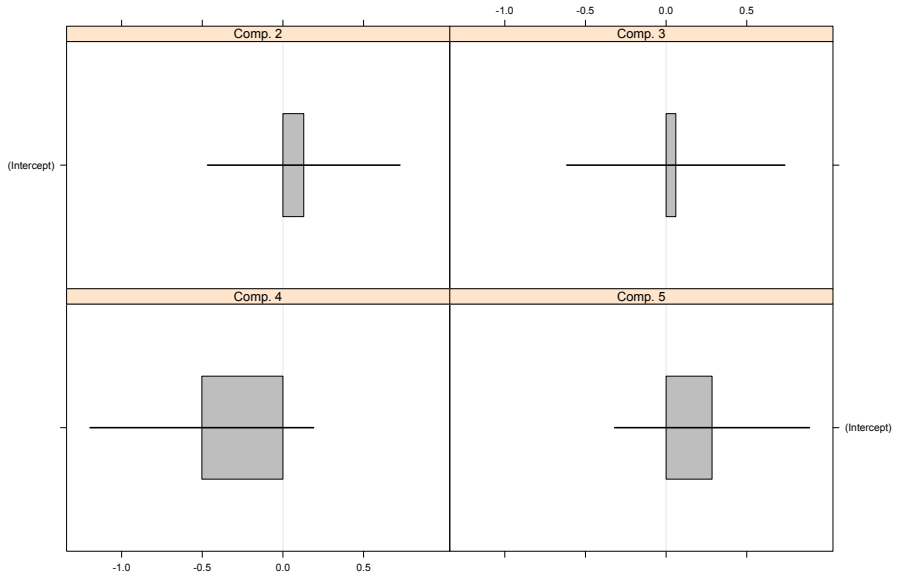


FIGURE E.5 – Coefficients de régression estimés des poids des composantes, produits UC.

E.3 Famille de produits Index-Link

E.3.1 Boxplot des coefficients

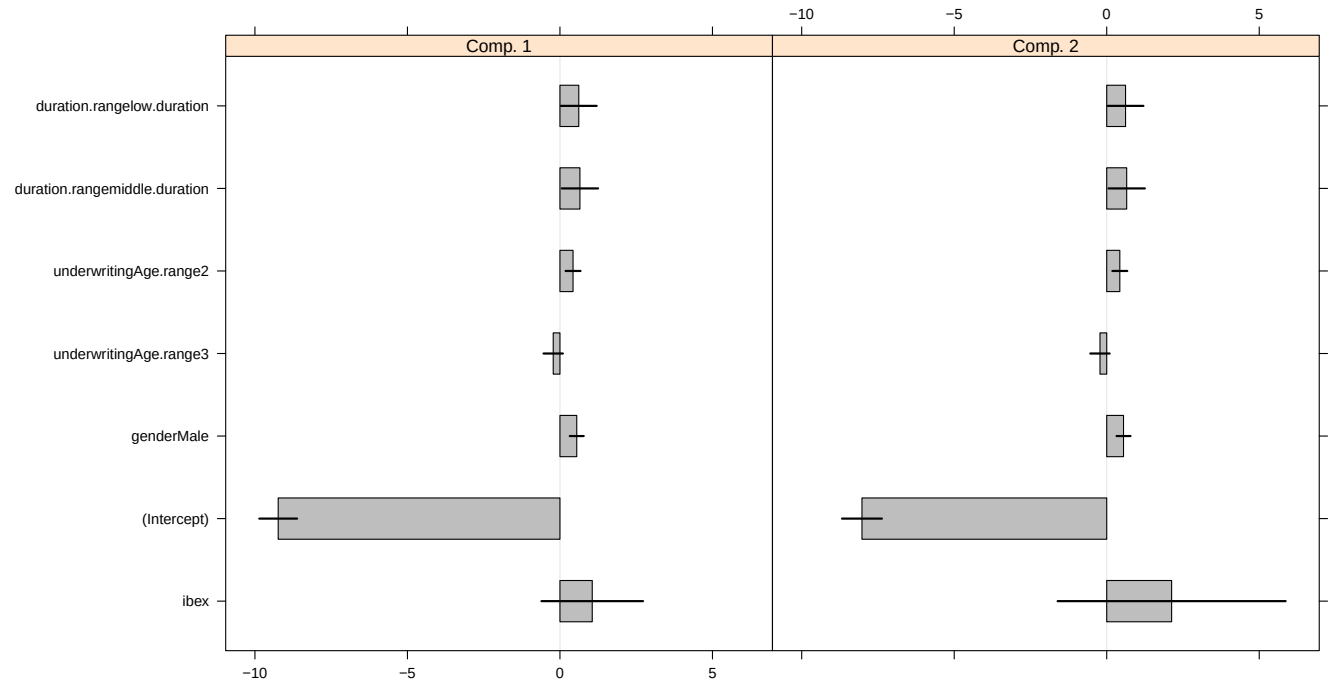


FIGURE E.6 – Coefficients de régression des composantes, produits Index-Link.

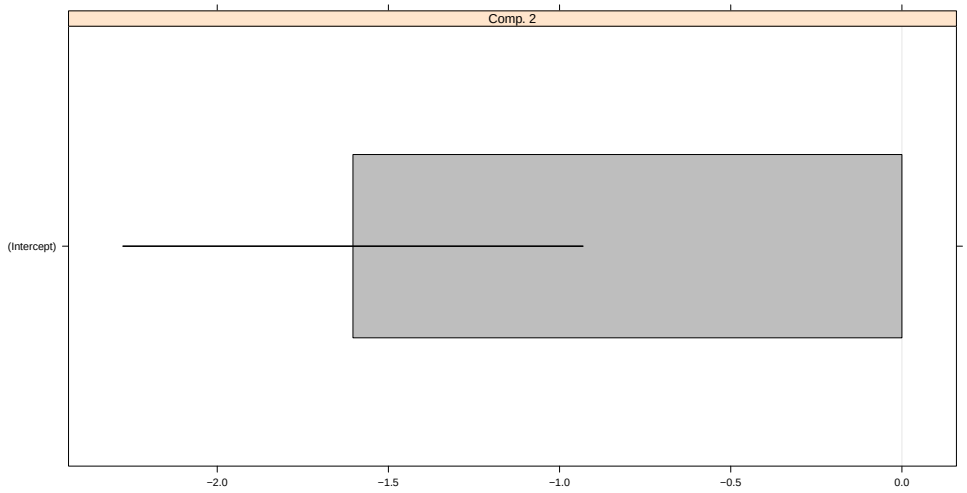


FIGURE E.7 – Coefficients de régression estimés des poids des composantes, produits Index-Link.

E.4 Famille de produits Universal Savings

E.4.1 Boxplot des coefficients

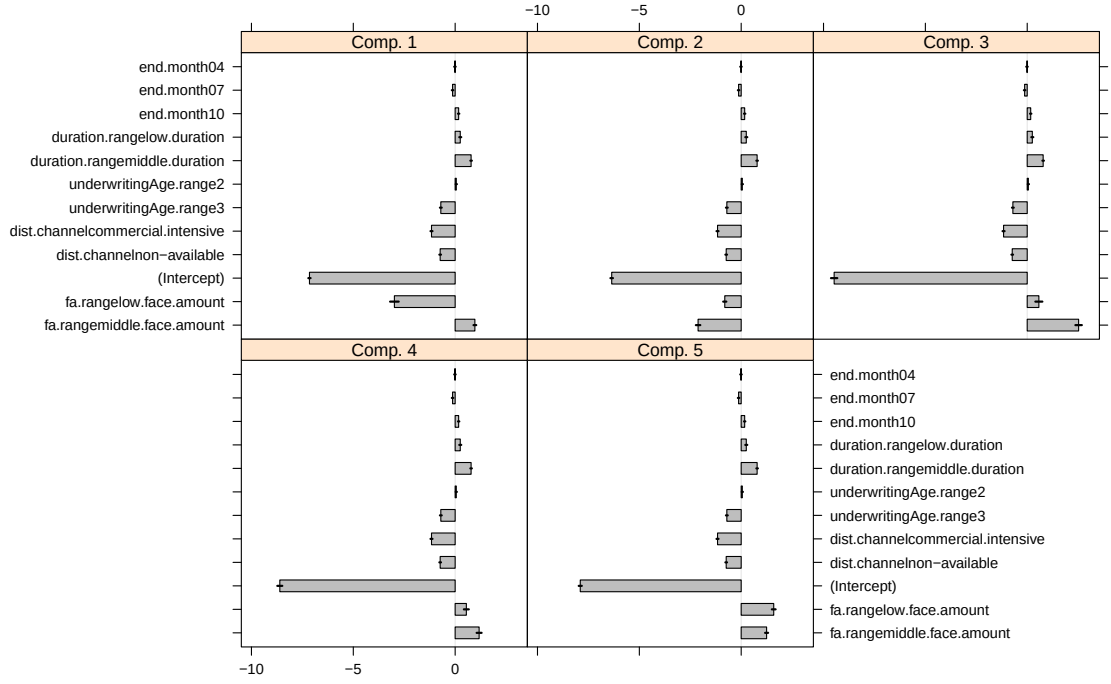


FIGURE E.8 – Coefficients de régression des composantes, produits Universal Savings.

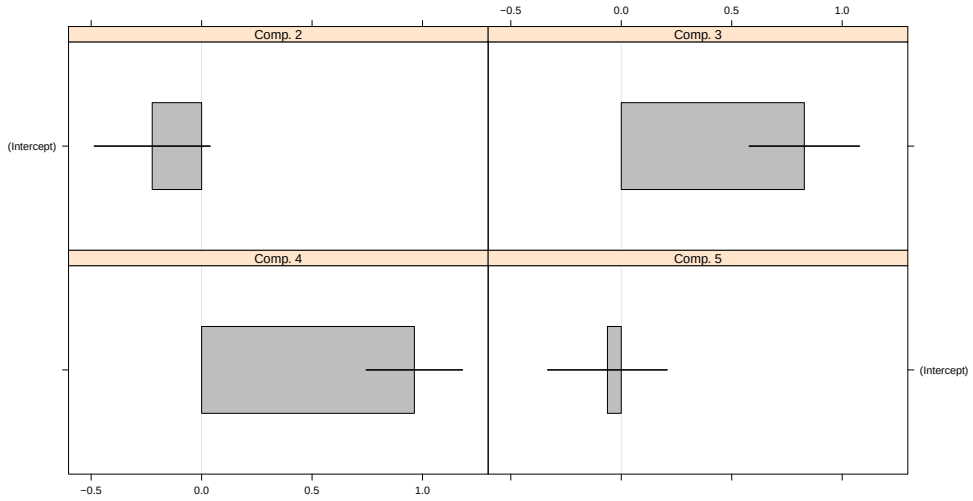


FIGURE E.9 – Coefficients de régression estimés des poids des composantes, produits Universal Savings.

E.5 Famille de produits Pure Savings

E.5.1 Boxplot des coefficients

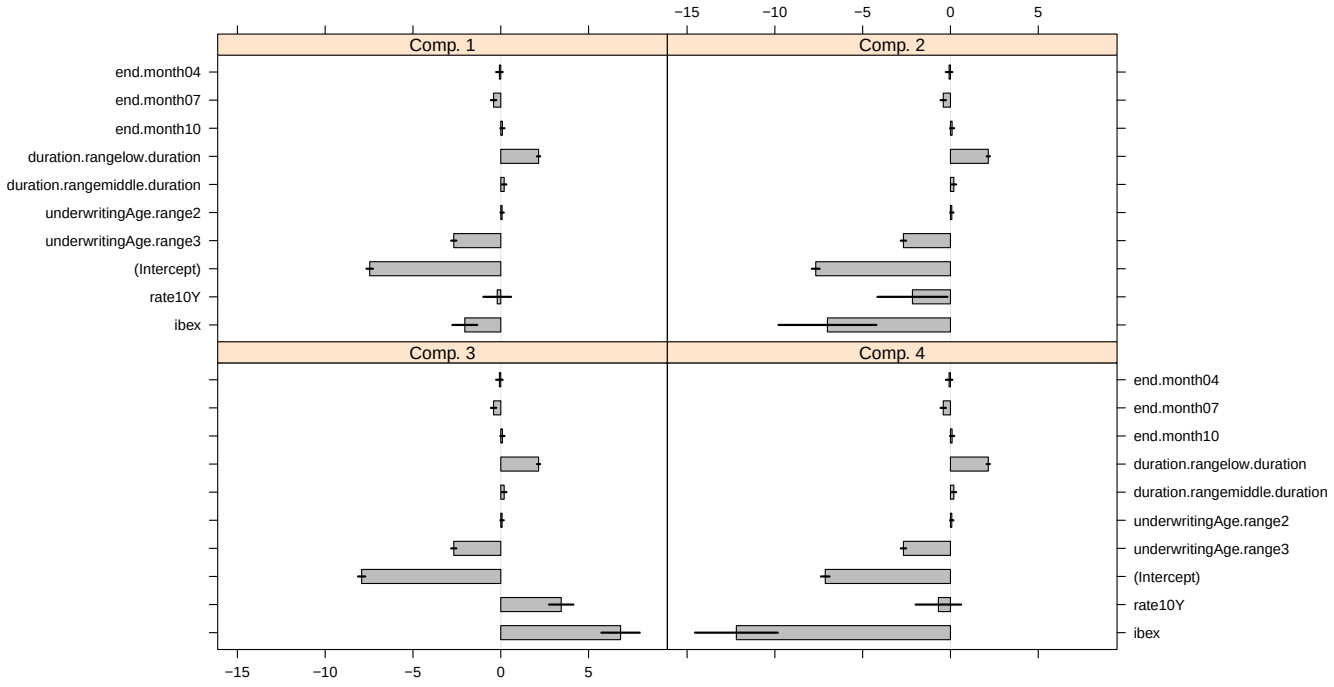


FIGURE E.10 – Coefficients de régression des composantes, produits Pure Savings.

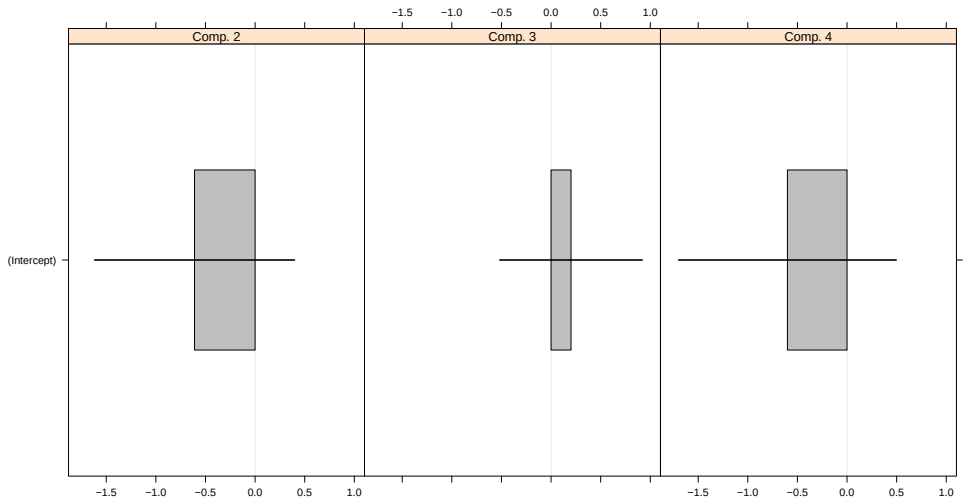


FIGURE E.11 – Coefficients de régression estimés des poids des composantes, produits Pure Savings.

E.6 Famille de produits "Structured Products"

E.6.1 Boxplot des coefficients

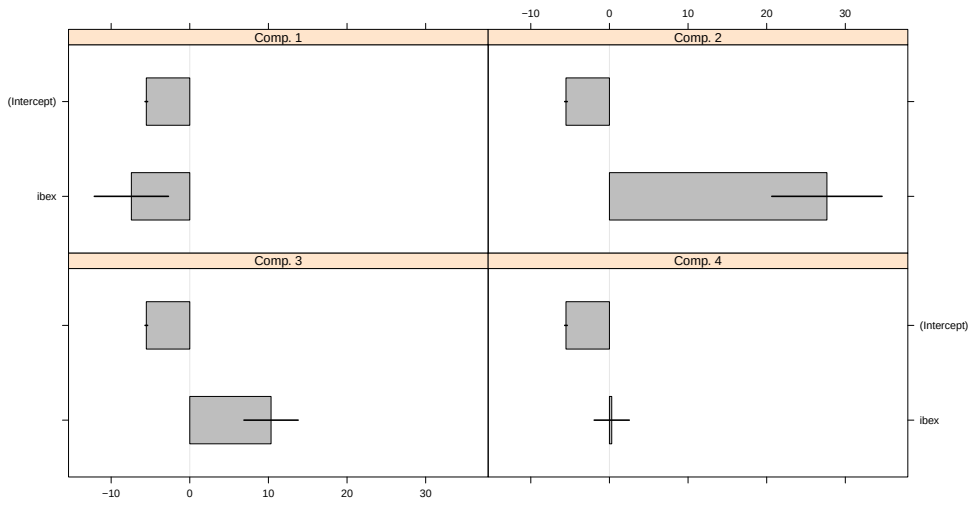


FIGURE E.12 – Coefficients de régression des composantes, produits structurés.

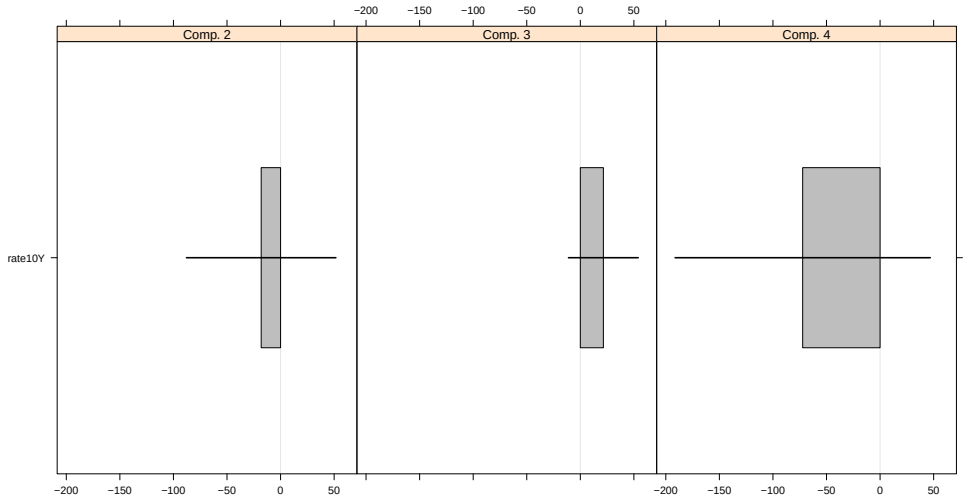


FIGURE E.13 – Coefficients de régression estimés des poids des composantes, produits structurés.