Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Léana Massounga Moukagni

Titre : Modélisation des mouvements d'arbitrage sur un portefeuille épargne suite à une incitation à arbitrer

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

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

Membres présents du jury de l'Institut des Actuaires

Signature

. . . . . .

. . . . . .

. . . . . .

Membres présents du jury de l'ISFA

. . . . . .

. . . . . .

. . . . . .

Entreprise :
Nom : Generali France

Signature :

Directeur de mémoire en entreprise :
Nom : Sandra Quinol

Signature :

Invité :
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Mémoire :
Modélisation des mouvements d'arbitrage sur un portefeuille épargne suite à une incitation à arbitrer

Auteure :
Léana Massounga Moukagni

Encadrants :
Sandra Quinol
Denys Pommeret

Master 2 - ACTUARIAT
Année universitaire 2019 - 2020

Remerciements

Avant tout développement sur cette étude, il apparaît opportun de commencer ce mémoire par des remerciements, à ceux qui m'ont beaucoup appris, et à ceux qui m'ont motivée à aller jusqu'au bout.

Aussi, je tiens tout d'abord à remercier Mme Sandra QUINOL, ma tutrice en entreprise ainsi que M. Denys POMMERET, mon tuteur universitaire, pour m'avoir encadrée tout en me faisant bénéficier de leur expérience.

Je voudrais également remercier mes proches, pour le courage, pour la force, et pour la motivation qu'ils m'ont apportés pendant la réalisation de ce mémoire.

Enfin je remercie tous les membres du service rentabilité et suivi du portefeuille de Generali, en particulier M. Arnaud Lemaire, pour les connaissances qu'ils ont partagées avec moi.

Résumé

Face aux exigences réglementaires et au contexte financier actuels, les assureurs redoublent de vigilance quant à leur activité d'épargne. Pour une gestion optimale de leurs portefeuilles, ils doivent suivre l'évolution de la répartition de l'encours entre les différents supports. L'ordre du jour est de faire basculer progressivement les assurés sur les supports en unités de compte, plus avantageux pour l'assureur. C'est dans cette optique que Generali a lancé en 2018 et en 2019 deux campagnes d'arbitrage sur l'un de ses portefeuilles. Ces campagnes ont pour but d'inciter les clients à acter des mouvements d'arbitrage des fonds en euros vers les fonds en unités de compte.

La présente étude propose la modélisation des mouvements d'arbitrage sur ce portefeuille à l'aide de modèles linéaires généralisés et de deux méthodes de *machine learning* : les arbres CART et les forêts aléatoires. Ensuite, dans la mesure où le pilotage de l'activité passe également par la connaissance client, il sera question de dégager les profils des assurés qui réalisent majoritairement des arbitrages des fonds en euros vers les fonds en unités de compte. Les profils des assurés mis en valeur pourront notamment servir d'indicateurs pour de futures actions commerciales plus ciblées.

Mots clés : Épargne, Arbitrage, Modèle linéaire généralisé, *Machine learning*, Profil des clients.

Abstract

Considering the regulatory requirements and the economic context, insurers are more vigilant about their savings business lines. For optimal portfolio management, they must find ways to divert policyholders from euro funds. With this aim in mind, Generali conducted two switching periods in 2018 and 2019 on one of its business lines.

The purpose of these switching campaigns was to encourage policyholders to make switch options from euro funds to other riskier supports. This thesis proposes the modelling of switching options on this business line using generalised linear models and two machine learning methods : CART trees and random forests. Then, as the management of the activity also requires policyholder knowledge, the profiles of the policyholders who mostly make switch options from the euro funds to to other riskier supports will be identified. The policyholder profiles highlighted could serve as indicators for more targeted marketing actions in the future.

Key words : Savings, Switch option, Generalised linear model, Machine learning, Policyholder profiles.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'assurance-vie compte parmi les placements préférés des français. Un contrat d'assurance-vie lie un assureur et son client. L'assureur s'engage, en contrepartie du paiement de primes, à verser une rente ou un capital à l'assuré ou à ses bénéficiaires à la fin du contrat. Ce type de contrat permet aux assurés de constituer un capital sur le long terme ou encore d'avoir un complément de revenus pour la retraite. De par sa fiscalité avantageuse, elle est très attractive pour les assurés.

Lorsque l'assuré réalise un versement sur son contrat, il a la possibilité d'investir son capital sur différents types de supports dont les fonds en euros et les fonds en unités de compte. Les fonds en euros s'adressent aux épargnants qui cherchent la sécurité. En effet, le capital investi sur ce type de fonds est garanti. Le client n'est exposé à aucun risque de perte pour ce type de supports. Les fonds en unités de compte, quant à eux, intéressent les clients aptes à prendre des risques. Sur ce type de fonds, le capital n'est plus libellé en euros mais en parts, dont la valeur fluctue en fonction des marchés boursiers. Il y a donc un réel risque de perte pour l'assuré sur ce type de supports. Ainsi, les fonds en euros ont tendance à être beaucoup plus attrayants pour les assurés que les fonds en unités de compte. Du point de vue de l'assureur, l'enjeu est tout autre.

Les contextes économiques et réglementaires actuels ne sont pas favorables pour les assureurs. Les marchés boursiers affichent des taux de plus en plus bas. Cette tendance à la baisse menace les placements à l'actif de l'assureur. D'après le contexte prudentiel de la norme Solvabilité II, les capitaux à immobiliser sont pilotés par les engagements de l'assureur mais aussi par sa gestion d'actifs. Dans le cas des fonds en euros, du fait de la garantie de capital, le risque est élevé pour l'assureur. Afin de gérer de façon optimale leurs portefeuilles épargne, les assureurs doivent être vigilants. Il devient essentiel pour eux de trouver des stratagèmes pour orienter davantage les assurés vers les fonds en unités de compte.

Dans la mesure où ils permettent de modifier la composition du portefeuille, les mouvements d'arbitrage réalisés par les clients sur leur contrat d'épargne représentent un levier important pour l'assureur. Conscient de cela, Generali a décidé de mettre en place pendant les années 2018 et 2019 des campagnes de *business transformation* sur l'un de ses portefeuilles épargne. Pendant ces campagnes, les assurés bénéficient de frais préférentiels afin de réaliser à moindre coût des arbitrages des fonds en euros vers les fonds en unités de compte. Ces campagnes ont pour but de préserver les fonds en euros ainsi que leurs rendements, et, de redresser la composition du portefeuille de Generali à son avantage.

L'objet de ce mémoire est de modéliser les mouvements d'arbitrage réalisés par les assurés de ce portefeuille. Dans un premier temps, il sera question de présenter le cadre dans lequel s'inscrit cette étude ainsi que les fondamentaux des contrats d'épargne. Ensuite, après une étude des éléments qui influencent les mouvements d'arbitrage, il sera question de modéliser ces derniers à l'aide de différents types de modèles : des modèles linéaires généralisés (GLM), des arbres CART (*classification and regression trees*) et des forêts

aléatoires. Chacun de ces modèles sera calibré sur une même base d'apprentissage. Puis, tous les modèles seront comparés entre eux au moyen des résultats obtenus sur une base de test. Pour terminer, les profils des assurés ayant une forte appétence pour les arbitrages des fonds en euros vers les fonds en unité de compte seront présentés. La connaissance client est cruciale pour le pilotage du portefeuille de l'assureur. Les assurés intéressés par les arbitrages des fonds en euros vers les fonds en unités de compte pourront notamment être le sujet d'autres politiques commerciales plus ciblées.

1. Mise en place de l'étude

Le but de cette partie est de présenter le cadre dans lequel s'inscrit cette étude et les différents mouvements qui rythment la vie d'un contrat d'épargne. Un premier état des lieux en matière d'arbitrages sera réalisé pour constater les résultats des campagnes d'arbitrages.

1.1 Les différents réseaux de distribution de Generali

Implantée dans l'Hexagone depuis 1832, Generali France a accéléré son développement à partir des années 1990, grâce à sa forte dynamique interne et à une politique de croissance externe. Avec plus de 11 milliards de chiffre d'affaire, environ 7600 collaborateurs et plus de 7 millions d'assurés, Generali France est devenu un acteur de référence en France notamment sur le marché de l'épargne.

Pour continuer sa conquête de la France, Generali peut compter sur son immense réseau de distribution composé :

— D'agents généraux :

L'agent est l'intermédiaire exclusif entre la compagnie et les clients. En tant que chef d'entreprise, il porte plusieurs casquettes. Il est à la fois commercial, gestionnaire, manager et dirige une équipe de collaborateurs. L'agent est détenteur de son portefeuille de clients. Il est rémunéré non seulement sur l'acquisition de nouveaux clients mais aussi sur la gestion et la valorisation de son portefeuille. La responsabilité de la compagnie est engagée en cas de faute de l'agent. Les missions d'un agent sont nombreuses. Il est chargé de conseiller les clients, de suivre leurs besoins, de les accompagner et de gérer son entreprise. Les agents généraux Generali sont au nombre de 800 et exercent leurs missions dans 950 points de vente de tailles plutôt importantes.

— De courtiers :

Le courtier est aussi un intermédiaire entre la compagnie et les clients. Contrairement à l'agent, c'est un professionnel indépendant qui représente ses clients auprès des compagnies d'assurance. Il a pour mission de trouver le meilleur produit au meilleur prix et de sélectionner le contrat le plus avantageux pour ses clients. Il accompagne ainsi le client dans le choix du contrat le plus profitable. On dénombre en France 18500 cabinets de courtage. Generali est le 2me fournisseur national des courtiers et réalise 21% de son chiffre d'affaires avec ce réseau.

— D'un réseau salarié :

Il est constitué de salariés de Generali. Ces derniers rencontrent principalement leurs clients / prospects à domicile ou sur leur lieu de travail. Le réseau salarié est orienté sur la clientèle de particuliers. Il propose des solutions d'assurance variées, notamment avec des garanties épargne, prévoyance et dommages. Son but est de suivre et développer son portefeuille aussi bien en nombre qu'en valeur. Il évalue et analyse régulièrement les besoins de ses clients afin de les satisfaire au mieux.

— De la direction des clients patrimoniaux :

Elle comporte 5 canaux de distribution :

- Les conseillers en gestion de patrimoine (CGPI) : ce sont des indépendants mandatés par leurs clients pour des missions diverses (stratégie patrimoniale, transmission,...).
- Les plateformes : ce sont des structures indépendantes ou des filiales du groupe. Elles mettent à disposition des CGPI une sélection d'offres en assurance-vie, défiscalisation, et une palette de services commerciaux et administratifs.
- Les banques privées : ce sont des filiales de banques nationales ou indépendantes positionnées sur le segment de clientèle fortunée. Elles procurent des conseils sur-mesure à leurs clients.
- Les banques régionales : ce sont des établissements positionnés entre les banques privées et les CGPI. Elles couvrent une région ou une localité. Elles effectuent toutes les opérations de banque dans leurs circonscriptions.
- Internet : Ce segment offre des services en ligne dans les domaines de la banque, du courtage et de l'information financière.

— Du réseau LFAC :

Historiquement, le réseau La France Assurance Courtage (LFAC) a un statut de courtage. Au premier Janvier 2009, LFAC est absorbée au sein de Generali Vie dans le cadre d'une transmission universelle de Patrimoine. Le personnel devient salarié de Generali Vie et le réseau prend la dénomination « La France Assurance Conseil ». Aujourd'hui, c'est un réseau en pleine transformation. Il a 3 sources d'activités principales : les institutions de prévoyance, les experts comptables et le contact direct. Il est spécialisé sur le marché de l'entreprise et des professionnels indépendants et commercialise principalement des produits de prévoyance et de retraite.

Generali propose ses produits à ses partenaires (apporteurs) qui les commercialisent ensuite auprès des clients. Le fonctionnement est le suivant (figure 1.1) :

FIGURE 1.1 – Mécanisme de distribution des produits de Generali

Grâce à sa panoplie de distributeurs, Generali peut assurer sa croissance. C'est un assureur très diversifié. Il propose des produits et des services qui couvrent tous les besoins : assurance dommages, épargne,

retraite et protection sociale pour les particuliers, les entreprises ou encore les professionnels. Generali est un assureur plutôt porté sur l'assurance-vie comme le montre la figure 1.2 de la page 15. D'ailleurs, un peu plus de la moitié de son activité repose sur l'épargne (Figure1.2).

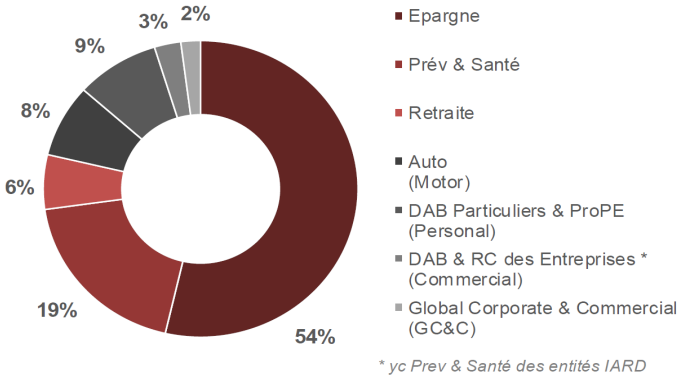

FIGURE 1.2 – Répartition de l'activité de Generali France - 2019

Le développement de l'activité d'épargne de Generali passe par le segment internet. Generali est le premier assureur sur le marché Internet de l'épargne. Ceci est possible grâce à la commercialisation de produits par des banques comme Boursorama et ING. La force du réseau internet réside dans la volonté de développer le digital afin de répondre aux nouvelles exigences et attentes des clients en terme d'accessibilité et d'efficacité. Pourtant, d'autres réseaux restent prometteurs. C'est notamment le cas du réseau salarié.

1.2 Les différentes offres produits de Generali

Le principe de l'assurance est simple : l'assureur s'engage en cas de survenance d'un risque à réaliser une prestation envers l'assuré moyennant le paiement d'une prime. Afin de couvrir les besoins de ses assurés et de les protéger au mieux, Generali leur propose un large panel de produits :

1.2.1 Les assurances de biens et responsabilité

La famille des assurances IARD (Incendies Accidents et Risques Divers), communément appelée assurances de biens, couvre principalement les risques subis par les biens matériels de l'assuré tels que les locaux, les équipements, les véhicules et les stocks. Cette grande famille englobe notamment :

— L'assurance automobile :

Ici, l'objet assuré est le véhicule. Une assurance automobile garantit une protection en cas d'accident de la route. L'assureur s'engage à verser une indemnisation aussi bien aux victimes des accidents qu'à ses assurés. Une assurance automobile comprend plusieurs garanties, notamment les garanties responsabilité civile (qui intervient dans le cadre de l'indemnisation des victimes hors conducteur si l'assuré est responsable) et dommages (pour les diverses réparations ou éventuels remplacements du véhicule).

— L'assurance MRH (multi risques habitation) :

  Il s'agit ici de protéger le logement et l'ensemble de ses occupants en cas de survenance de l'un des évènements suivants : incendie, vol, dégât des eaux, catastrophes naturelles ou évènements climatiques. Les assurances MRH incluent la garantie responsabilité civile « Chef de famille » qui couvrent la famille en cas de faute ou de dommages causés à des tiers dans le cadre de la vie privée.

— L'assurance MRC (multi risques commerciale) :

  Elle a pour but d'assurer l'entreprise ainsi que le local professionnel contre les accidents liés à l'activité professionnelle. Elle permet à l'entreprise de rapidement reprendre son activité en cas de sinistre.

1.2.2 Les assurances de personnes

Les assurances de personnes, par opposition aux assurances de biens, ont pour objet de couvrir (collectivement ou individuellement) la personne humaine. Elles couvrent les risques portant atteinte à l'intégrité physique de la personne. Elles protègent les personnes contre les accidents corporels, la maladie, l'incapacité, l'invalidité et le décès.

— L'assurance santé :

  Elle a été mise en place pour financer les dépenses de santé des assurés en complément du régime de base de la sécurité sociale. Elle prend en charge différents types de soins à l'instar de l'hospitalisation, des consultations médicales, des médicaments, des soins dentaires.

— L'assurance prévoyance :

  Elle permet de financer la perte de revenus en cas d'aléa portant sur la personne. Elle couvre notamment les risques de décès, d'incapacité et d'invalidité. Elle intervient en complément du régime de base de la sécurité sociale pour le remboursement d'indemnités journalières pour l'incapacité et la maternité, pour le versement de rentes pour l'invalidité, le versement d'un capital ou d'une rente aux ayants droits en cas de décès.

— L'assurance emprunteur :

  Elle fait office de sécurité pour garantir un emprunt. Elle prend le relais en cas de décès, d'invalidité ou d'incapacité ou de perte d'emploi pour rembourser l'établissement de crédit. Depuis l'entrée en vigueur le 1er Janvier 2015 de la loi Hamon, les assurés ont la possibilité de changer d'assurance emprunteur

dans les 12 mois qui suivent la signature du contrat.

— La GAV (Garantie accidents de la vie) :

Elle finance les conséquences pécuniaires des accidents de la vie courante entraînant le décès ou des séquelles à vie. Dans le cadre du décès, elle prévoit le versement à la famille de l'assuré d'un capital correspondant au revenu réel de la personne décédée, ainsi que d'une somme correspondant au « préjudice d'affection ». Si l'assuré est atteint de séquelles définitives, elle prévoit le versement d'un capital correspondant à la perte de l'intégrité physique, à la perte réelle de revenu, et à des frais nécessaires à l'aménagement de la vie de la personne (travaux au domicile, aide d'une tierce personne).

Ci-dessous un panorama des risques couverts par les assurances de personnes :

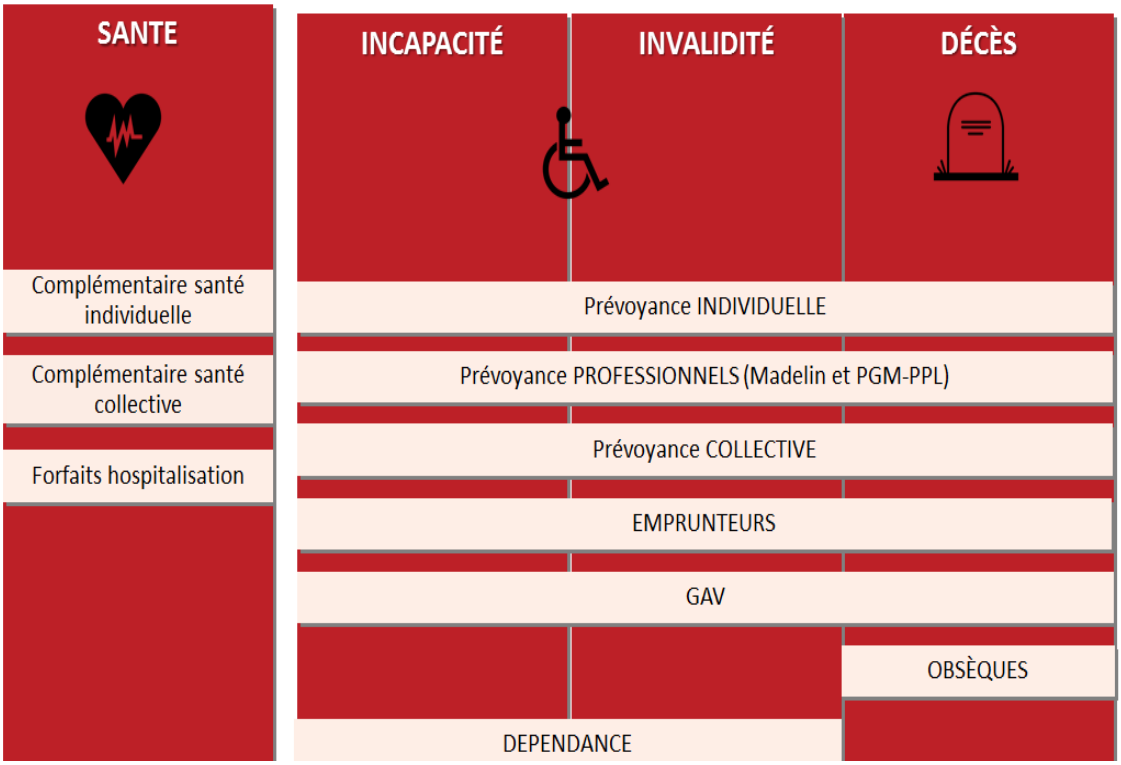

FIGURE 1.3 – Panorama des assurances de personnes

1.2.3 La retraite

En France, le système de retraite est organisé en 3 piliers :

— Le pilier 1 : Il est composé des régimes légaux obligatoires, à savoir le régime de base de la Sécurité Sociale et le régime complémentaire obligatoire de AGIRC - ARRCO [1]. Tous les actifs sont obligés d'y cotiser. Ils varient en fonction du statut professionnel et/ou du métier de l'assuré ;

— Le pilier 2 : Il comprend les régimes collectifs d'entreprise à adhésion facultative ou conventionnellement obligatoire. Les pensions de ce régime viennent souvent compléter les pensions des régimes du premier pilier jugées insuffisantes ;

— Le pilier 3 : Il comporte les régimes supplémentaires à adhésion facultative. Il s'agit de produits réglementés d'épargne (individuels ou collectifs) dédiés à la retraite.

La représentation pyramidale ci-dessous permet d'avoir une vision globale du système de retraite français :

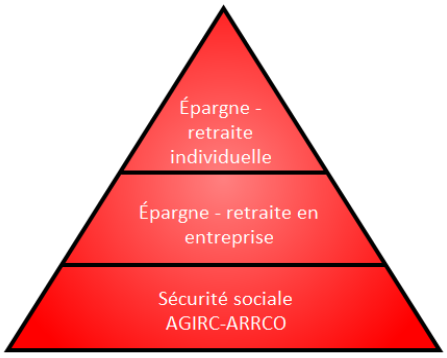

FIGURE 1.4 – Représentation du système de retraite en France

Chaque pilier donne droit à une pension. La pension totale perçue par un individu est obtenue en faisant la somme des différentes pensions.

Les produits de retraite proposés par Generali rentrent dans le cadre des piliers 2 et 3. Il s'agit de rentes viagères versées à partir de la date de départ à la retraite. Ces rentes peuvent être réversibles (versées au conjoint en cas de décès de l'assuré), à annuités garanties. Le montant de la rente versée dépend de plusieurs paramètres dont l'âge de l'assuré (lors du départ à la retraite), le montant total des cotisations versées, les options de rente choisies, le taux technique garanti dès l'adhésion et la table de mortalité utilisée.

1. Association générale des institutions de retraite complémentaire des cadres - Association pour le régime de retraite complémentaire des salariés.

1.2.4 L'épargne

Les contrats d'assurance-vie comptent parmi les placements préférés des français. Ils sont attractifs car ils bénéficient d'une fiscalité avantageuse et de rendements en moyenne assez élevés. Lors de la souscription d'un contrat d'épargne, Generali offre la possibilité à ses clients de choisir entre trois types de supports :

— Les fonds en euro :

Le fonctionnement d'un fonds euro est similaire à celui d'un livret d'épargne classique. Le capital versé par l'assuré est revalorisé chaque année à un taux de participation aux bénéfices déterminé par l'assureur. Ainsi, le capital versé par l'assuré est garanti. Il n'y a aucun risque de perte pour le client. Ce type de fonds repose généralement sur des actifs sécurisés. Ici, Generali porte tous les risques mais s'assure de proposer à ses clients un large choix de fonds euro. Les rendements de ces derniers sont tous différents et varient en fonction des actifs sur lesquels les capitaux sont investis. Cependant, la tendance est à la baisse pour les rendements des fonds euro depuis plusieurs années déjà, comme en témoigne la figure 1.5 de la page 19.

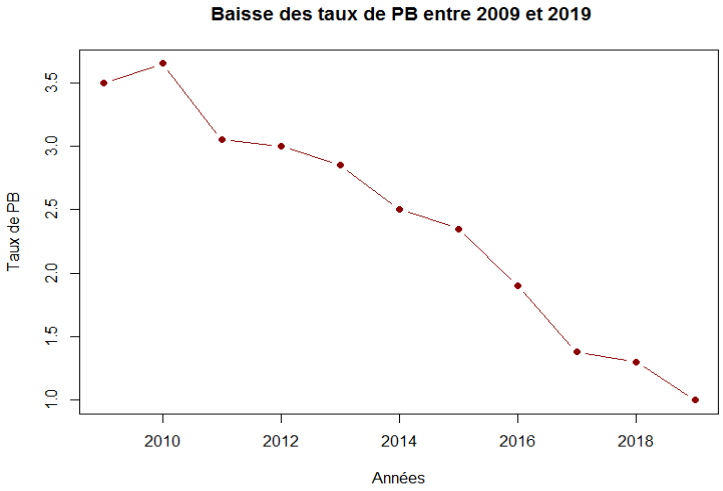

FIGURE 1.5 – Taux de PB sur un fonds euro entre 2009 et 2019

— Les fonds en unités de compte (UC) :

Pour les fonds en UC, le mécanisme diffère. L'investissement du client pour ces fonds se mesure en nombre de parts. Generali se porte garant uniquement du nombre de parts détenues mais pas de leur valeur. Il y a des risques de perte de capital. Contrairement aux fonds euro, le client porte tous les risques. Les fonds UC reposent sur des actifs très dynamiques. En moyenne, ils offrent des rendements plus intéressants que les fonds euro. Il y a deux types de fonds UC : les fonds UC gérés par Generali dits UC maison et les autres fonds UC dits fonds UC tiers gérés en externe. Ces fonds ont été mis en place

compte tenu du contexte économique (avec les taux à la baisse) afin de ne pas diluer les rendements des fonds en euros.

— Les fonds euro croissance :

Arrivé en 2014, ce type de fonds se trouve à mi-chemin entre les fonds euro et les fonds UC. Il présente l'avantage d'être moins risqué que les fonds UC et plus rentable que les fonds euro. Le capital investi sur ce type de support est divisé en deux : une première partie en euros qui fait office de provision mathématique, et, une autre partie exprimée en parts de diversification. La provision mathématique est garantie en totalité ou partiellement à l'échéance du contrat (8 ans ou plus). Un retrait anticipé peut entraîner des pertes car dans ce cas le client ne bénéficie d'aucune garantie de capital. En 2019, le fonds euro croissance a été réformé par la loi Pacte puis mis à jour via un décret et un arrêté. Depuis lors, il existe une unique provision dite provision de diversification qui est exprimée directement en euros.

1.2.4.1 Les mouvements usuels :

Tout au long de la durée de leurs contrats, les clients ont la possibilité de faire plusieurs mouvements. La vie d'un contrat d'épargne est rythmée par des versements, des rachats et des arbitrages.

— **Les versements**

Les versements sont de deux types. Ils peuvent intervenir à n'importe quel moment du contrat. Dans ce cas, il s'agit de versements libres. Le montant des versements libres est à la convenance du client et peut différer d'un versement à l'autre. Les versements peuvent également arriver de façon périodique. Il s'agit dans ce cas de versements programmés. Ces derniers sont généralement tous d'un même montant décidé à l'avance par le client. Ils sont souscrits dans le cadre d'une option sur le contrat.

— **Les rachats**

Les contrats d'assurance-vie permettent à l'assuré de récupérer partiellement (dans ce cas, il s'agit de rachat partiel) ou totalement (rachat total) l'épargne capitalisée avant le terme du contrat. Lors des rachats, seule la plus-value générée par le contrat est taxée par l'Etat. Deux taxes sont appliquées : la taxe fiscale et la taxe sociale. La première relève du code des impôts, tandis que la seconde relève des lois de financement de la sécurité sociale.

La fiscalité appliquée à un contrat dépend de la date d'ouverture du contrat et des dates auxquelles l'assuré a effectué des versements. Elle évolue avec l'ancienneté du contrat. Plus le contrat est ancien, plus la fiscalité est avantageuse. La fiscalité appliquée change généralement au bout de 4 ans et 8 ans d'ancienneté.

Il existe 3 compartiments de fiscalité appelés : « C1 », « C2 » et « C3 ». Ils sont répartis de la façon suivante :

Date de versement des primes	Montant des primes	Fiscalité en vigueur
Jusqu'au 31/12/1982	×	Exonéré d'impôts
Entre le 01/01/1983 et le 25/09/1997	×	Compartiment C1
Entre le 26/09/1997 et le 31/12/1997	$< 30\ 490$	Compartiment C1
Entre le 26/09/1997 et le 31/12/1997	$\geq 30\ 490$	Compartiment C2
Entre le 01/01/1998 et le 27/09/2017	×	Compartiment C2
Entre le 28/09/2017 et aujourd'hui	×	Compartiment C3

Table 1.1 – Fiscalité en vigueur en fonction de la date de versement des primes

— **Les arbitrages**

Un arbitrage est une opération qui consiste à modifier la répartition de la valeur atteinte entre les différents supports financiers du contrat. À la souscription, le client a la possibilité de choisir une option de gestion en fonction de son profil de risque (dans ce cas, le contrat est en gestion pilotée) ou de suivre l'évolution de son contrat lui-même (contrat en gestion libre). Les deux options de gestion sont exclusives l'une de l'autre. Ainsi, il existe deux types d'arbitrages : les arbitrages libres et les arbitrages programmés. Les premiers sont expressément demandés par les clients, tandis que les seconds interviennent dans le cadre de la gestion pilotée. Il s'agit d'arbitrages automatiquement faits sur le contrat.

Tous les arbitrages supportent des frais. Ces frais sont proportionnels au montant arbitré et ne peuvent excéder un plafond pour certains types de contrats. Les arbitrages sont éventuellement assortis de restrictions définies dans les conditions générales du contrat. Il peut s'agir par exemple de l'éligibilité de certains supports d'arrivée pour les arbitrages. Le but est de protéger au maximum les fonds euro afin de limiter les engagements de l'assureur. La plupart du temps, les assurés ont la possibilité de faire des arbitrages entre tous les types de fonds.

1.2.4.2 Les options

En dehors de l'option de rachats partiels programmés et de l'option de gestion mentionnée plus haut, d'autres options peuvent être souscrites dans le cadre d'un contrat d'épargne multi-supports. Il s'agit notamment de :

— La sécurisation des plus-value :

  Cette option est mise en place afin de permettre au client de « sécuriser » la plus-value dégagée sur les fonds UC disponibles sur le contrat. Lorsque le client choisit l'option sécurisation des plus-value, Generali lui propose de transférer de façon automatique, à partir d'un seuil déterminé, la plus-value constatée, sur les supports en unités de compte sélectionnés vers un support de sécurisation.

— Les avances :

Une avance est un prêt consenti par l'assureur à l'assuré sur son contrat, sans que cela ne modifie l'épargne de ce dernier. C'est une alternative aux emprunts bancaires. Tant qu'elle n'est pas soldée, une avance fait courir des intérêts. Le taux d'intérêt est fixé par l'assureur.

— La limitation des moins-values :

Cette option est mise en place pour protéger le client contre une baisse en dessous d'un certain seuil de son épargne sur les supports en unités de compte. Dès lors que le seuil déterminé aura été constaté, l'assureur transfère totalement la valeur atteinte de chaque support sélectionné vers un support de sécurisation.

— La garantie plancher :

Cette option permet de garantir aux bénéficiaires, lors du décès de l'assuré, une prestation qui correspond au maximum entre l'épargne atteinte du contrat et le capital minimum déterminé à la souscription du contrat. Ce capital peut ensuite varier suivant les versements et retraits effectués sur le contrat.

1.2.5 La multi-détention du réseau salarié

Le réseau salarié a la particularité de proposer des offres personnalisées à ses clients. Ces dernières incluent de nombreuses garanties : épargne -retraite, prévoyance, MRH et auto. Ces offres sont faites sous la forme de multi-équipement.

Les packages sont généralement constitués d'un « socle » auquel viennent se rajouter d'autres options. Le socle correspond aux garanties épargne-retraite et les garanties complémentaires sont très souvent des garanties prévoyance (incapacité, invalidité), GAV, chomâge, garantie plancher, santé, garantie emprunteur et, plus rarement, auto ou MRH. Le client effectue les associations des différents produits à disposition en fonction de ses besoins et de ses objectifs. Les différentes associations correspondent à des catégories de clients. Il existe des offres ciblées pour les jeunes, les familles et les seniors.

Cette multi-détention a de nombreux atouts :

— Atout simplicité : Les règles de souscription sont allégées. Les clients ont une facilité d'accès à tous les produits. La gestion est plus facile pour eux : l'encaissement de prime est unique ;
— Atout sécurité : Les contrats tout en un donnent la possibilité au client de constituer, à son rythme, une épargne-retraite, tout en bénéficiant de performances d'actifs attractives. Ils permettent d'assurer le quotidien en cas de maladie ou d'accident, et de préparer l'avenir.

Tout au long de cette étude, il sera question de deux produits en particulier :

— **Generali Protection Vie** (GPV) :

Il s'agit d'un contrat d'assurance-vie individuel. Il offre au client la possibilité de se constituer soit un capital, soit une rente par le biais de cotisations périodiques et de versements libres. Ce produit s'adresse principalement aux familles installées et aux seniors. Ce contrat d'épargne multisupports bénéficie d'une offre financière très variée avec 27 fonds Generali Investments. Depuis son lancement en 2015, le portefeuille est en croissance permanente, même si le rythme des affaires nouvelles ralentit (Voir figure 1.6 ci-après). En 2016, le portefeuille comptait environ 38 000 contrats alors qu'en 2019, ce nombre a presque triplé jusqu'à atteindre 103 000. Les clients GPV détiennent en moyenne 2,9 contrats.

Ce produit est souvent associé à d'autres garanties prévoyance et à la GAV.

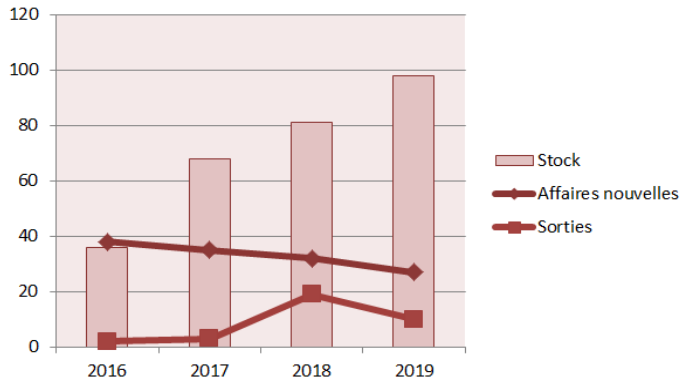

FIGURE 1.6 – Évolution de GPV en milliers de contrats

— **GPA Profil Epargne** (GPE) :

Plus ancien que GPV, il a été lancé en Juin 2004 dans le but de rétablir un équilibre avec les branches prévoyance et retraite. Contrairement à GPV, il s'agit d'un contrat d'épargne à prime unique. Il permet ainsi la constitution d'un capital ou d'une rente (par des versements libres ou programmés), ou alors le versement d'un capital en cas de décès au profit de bénéficiaires désignés. Cette souplesse, dans le cadre des versements, donne la possibilité au client d'épargner à son rythme. Divers supports d'investissements sont disponibles. Ce produit est généralement combiné avec des produits prévoyance et GAV au sein d'un même pack. Commercialisé avant GPV, il compte à la fin de 2019 un total de 250 000 contrats (Voir figure 1.7 de la page 24).

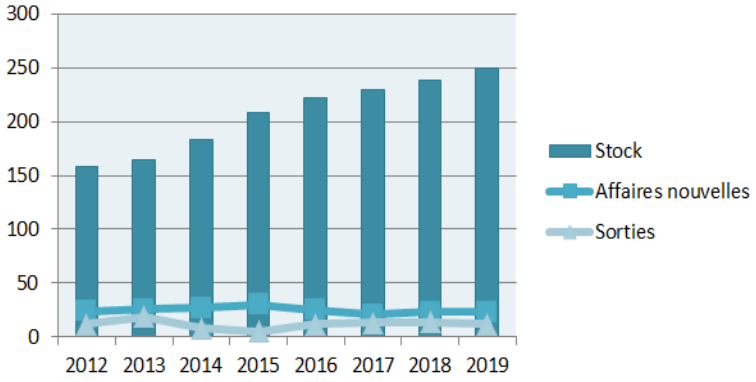

FIGURE 1.7 – Évolution de GPE en milliers de contrats

1.3 Les campagnes de *business transformation*

L'étude menée dans ce mémoire sera réalisée en utilisant les données des contrats d'épargne du réseau salarié. Il suscite un intérêt particulier car il a été sujet de deux campagnes d'arbitrage : une première de Mai à Octobre 2018, et une deuxième pendant les mois de Mai et Juin 2019 (voir figure 1.8).

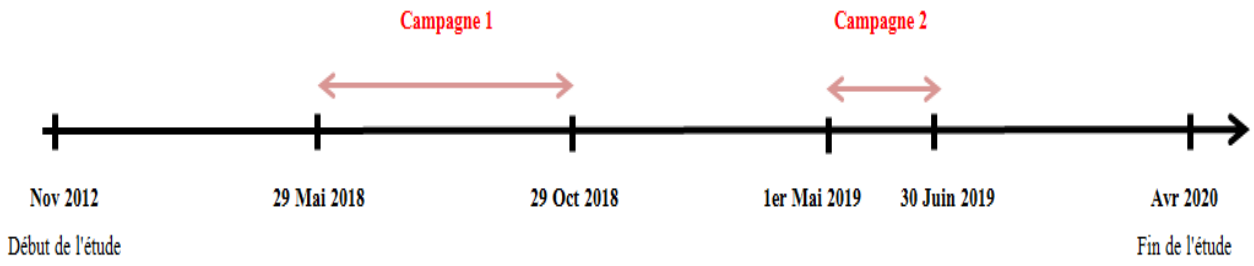

FIGURE 1.8 – Répartition de l'étude dans le temps

1.3.1 L'objectif des campagnes

Les clients du réseau salarié ont le choix entre les deux types de supports pour l'investissement de leur épargne. Ils ont à leur disposition un seul fonds euro et une large gamme de fonds UC (maison et tiers). Ainsi, ils peuvent réaliser plusieurs sortes d'arbitrages entre tous ces fonds. Seuls les arbitrages entre fonds euro sont impossibles dans la mesure où il n'y a qu'un seul fonds euro disponible. La figure 1.9 présente une synthèse des mouvements d'arbitrage possibles.

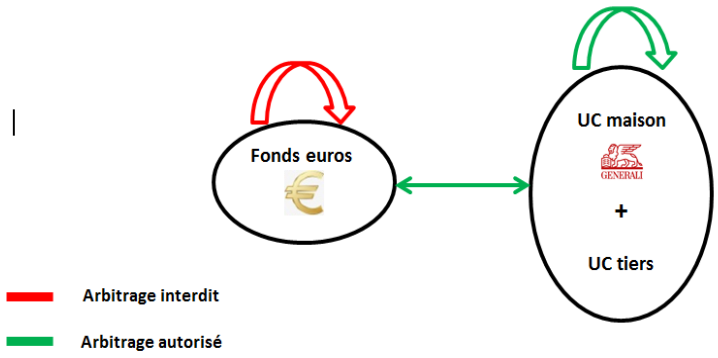

FIGURE 1.9 – Les mouvements d'arbitrage possibles

Les mouvements d'arbitrage sont généralement assortis de frais à la charge du client et d'une commission versée à l'apporteur d'affaire. Ces derniers sont exprimés en pourcentage du montant arbitré. Généralement, les frais d'arbitrage sont de 0,8%.

Malgré les nombreux fonds UC à leur disposition, les clients ont tendance à placer leur épargne majoritairement sur le fonds euro. Du fait de la garantie de capital, les engagements de l'assureur sont élevés pour ce type de fonds. Par ailleurs, lorsqu'il y a trop de clients sur un fonds euro, son rendement est dilué. Les taux servis sont moins attractifs. De plus, les rendements du fonds euro sont en baisse depuis plusieurs années. Un nombre de clients important sur un fonds euro dilue d'autant plus le rendement de ce dernier. Le défi pour l'assureur est de continuer de proposer des supports d'investissement alternatifs aux assurés. C'est dans cette optique que Generali a élargi sa gamme de fonds UC maison en 2018. La première campagne de *business transformation* s'inscrit dans ce cadre là. Elle a été mise en place afin d'inciter les clients à acter des mouvements d'arbitrage de tous les types de fonds vers les fonds UC maison. La deuxième période de *business transformation*, moins ciblée, encourageait les mouvements d'arbitrage du fonds euro vers les fonds UC en général (aussi bien UC maison que UC tiers).

Au regard de la directive sur la distribution d'assurance (DDA) [2] entrée en vigueur après la première campagne d'arbitrage, un conflit d'intérêt pourrait naître par rapport à l'application de taux de frais d'arbitrage et de commission différenciés en fonction du type d'arbitrage. Dès lors que le montant de rémunération relatif aux arbitrages est peu significatif au vue de la rémunération totale des conseillers commerciaux, ce conflit est écarté. De plus, des critères qualitatifs permettant le suivi du devoir de conseil du conseiller commercial vis-à-vis de son client ont été instaurés. Il s'agit notamment de la mise en place des "contrôles-compagnie" concernant le nombre d'arbitrages effectués par an sur un même contrat. Par ailleurs, il y a également la reprise des commissions calculées au prorata des commissions versées sur les 12 derniers mois.

Ainsi, pendant les périodes de *business transformation*, les frais d'arbitrage et de commissionnement ont été modifiés afin de rediriger au maximum les clients vers les fonds UC. Le taux de commissionnement a été augmenté pour les arbitrages du fonds euro vers les fonds UC. Les arbitrages entrant vers les fonds UC maison étaient gratuits (avec des taux de frais d'arbitrage de 0 %).

2. C'est une directive européenne en vigueur depuis 2016, dont l'ordonnance et les décrets ont été publiés en France en mai et juin 2018. La DDA vient en fait abroger la Directive sur l'Intermédiaire en Assurance (aussi appelée IMD1). Elle a été mise en place afin d'améliorer la réglementation sur le marché des assurances en vue d'instaurer une concurrence équitable entre les différents assureurs.

1.3.2 Les résultats des campagnes

Les campagnes d'arbitrage ont eu plusieurs impacts : sur la composition de l'encours du portefeuille, sur les montants et les volumes arbitrés.

☐ **Évolution de l'encours du portefeuille**

Depuis 2013, le portefeuille étudié a continuellement évolué. L'encours a grossi d'environ 38% entre Janvier 2013 et Janvier 2020 (voir figure 1.10). Néanmoins, la constitution du portefeuille reste globalement la même pendant cette évolution. L'encours est majoritairement placé sur le fonds euro, ensuite sur les fonds UC tiers. Les fonds UC maison ne représentent qu'une petite partie de l'encours total. Pour Janvier 2020 par exemple, le fonds euro compte pour plus de 66 % de l'encours total, suivi par les fonds UC tiers avec un pourcentage de 22 % ; les fonds UC maison comptent seulement pour 12%.

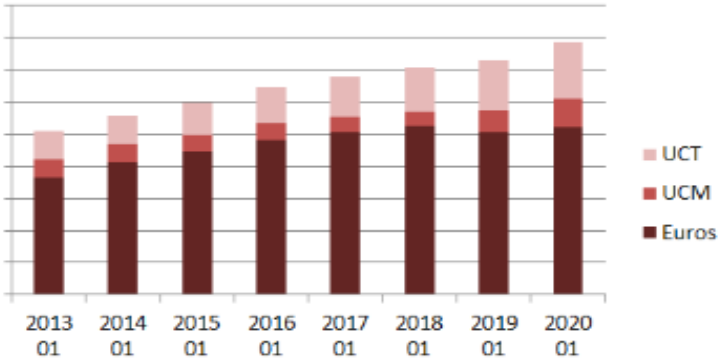

FIGURE 1.10 – Répartition de l'encours par type de fonds en montant

Après les campagnes de *business transformation*, la structure du portefeuille a légèrement été modifiée. La part de l'encours relative au fonds euro est plus élevée que celle des autres fonds. Sur les années avant les campagnes, elle est en moyenne de 74%. En 2019 et en 2020, elle chute légèrement et passe respectivement à 69 % et 66 %. La part d'UC maison est relativement faible. Elle décroît entre 2013 et 2018. Entre Janvier 2018 et Janvier 2020, elle double quasiment et passe de 6 % à 11 %. Les campagnes de *business transformation* ont également boosté la part d'UC tiers. Elle passe de 17 % en moyenne sur les années avant les campagnes d'arbitrages à 21 % en moyenne après. Ces informations sont présentées par la figure 1.11.

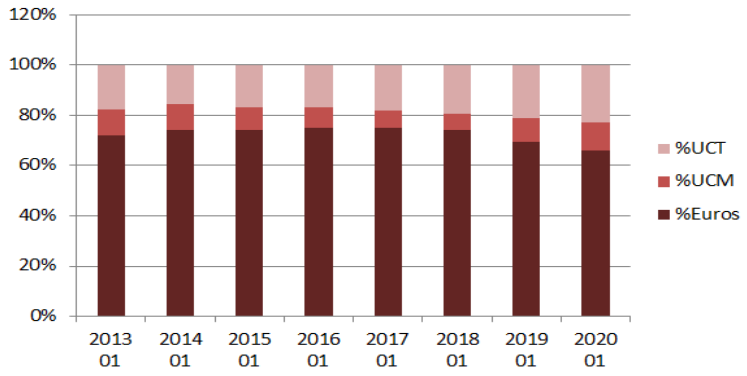

FIGURE 1.11 – Répartition de l'encours par type fonds en pourcentage

☐ Évolution du nombre d'arbitrages

En général, les assurés du réseau salarié n'effectuent pas beaucoup d'arbitrages. Le nombre d'arbitrage moyen par mois est seulement de 227, tandis que, le nombre de contrats étudiés entre 2012 et 2020 a augmenté passant de 158 000 contrats à environ 350 000 contrats. Néanmoins, entre Janvier et Mars 2017, une hausse du nombre d'arbitrages est constatée. Il y a une petite hausse sur la courbe du nombre d'arbitrages à cette période. Cela reste cependant négligeable devant le pic qu'il y a en 2018. Le nombre d'arbitrages effectués pendant la deuxième partie de l'année 2018 a explosé. Ceci est essentiellement dû à la première campagne d'arbitrages. La moyenne d'arbitrages observés pendant cette période est de 1950 avec un maximum atteint en Juillet 2018 pour un nombre de 3 103 arbitrages. La deuxième campagne a bien moins fonctionné que la première. Avec un nombre d'arbitrages moyen par mois de 495, elle se classe au même niveau que la hausse de 2017. Ce manque de réactivité de la part des assurés peut s'expliquer par le fait qu'elle dure moins longtemps que la première campagne (2 mois contre 5 mois). Après les deux campagnes, la moyenne mensuelle du nombre d'arbitrages chute à 252.

☐ Évolution du volume arbitré

Les campagnes d'arbitrages ont généré un nombre important d'arbitrages pour un montant total arbitré de 198 000 000 d'euros. Les arbitrages effectués pendant ces périodes sont majoritairement des arbitrages euros vers UC. Les clients ont désinvesti massivement le fonds euro au profit des fonds UC maison et tiers. Les campagnes ont été une réussite. De même que pour le nombre d'arbitrages, le volume arbitré pendant la première campagne est largement supérieur à celui arbitré pendant la seconde période. Les clients ont eu la réaction escomptée en arbitrant plus de ressources sur les fonds UC maison que sur les fonds UC tiers pendant la première campagne. Lors de la deuxième campagne, ils ont investi sur les fonds UC maison et UC tiers approximativement pour le même montant.

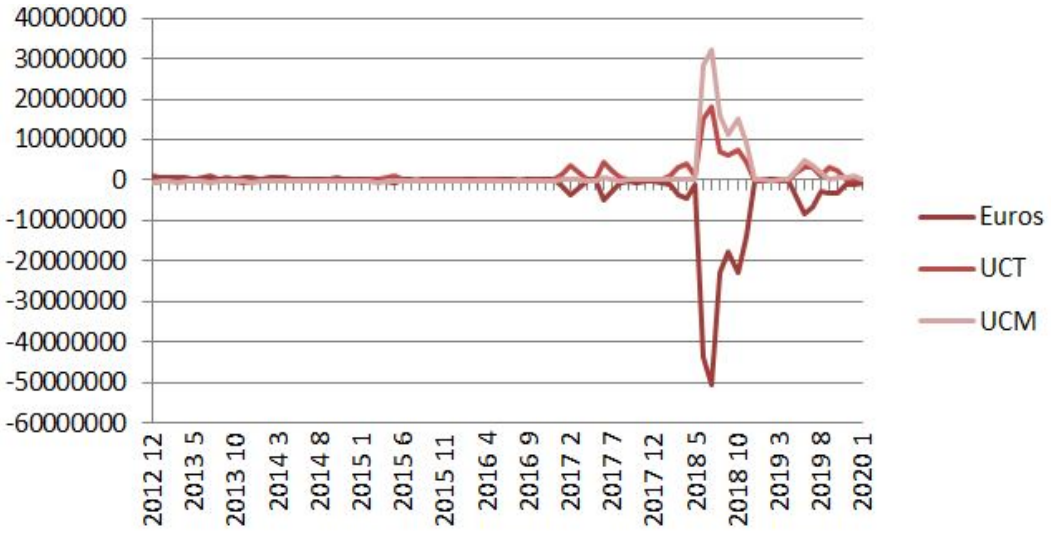

FIGURE 1.12 – Montant arbitré (en euros) par type de fonds entre Novembre 2012 et Janvier 2020

Les arbitrages actés avant et après les périodes de campagne sont écrasés par le volume important d'arbitrages qu'il y a pendant les campagnes. En regardant de plus près, des arbitrages se démarquent tout de même. Mais le comportement des assurés est inversé. Ils ont tendance ici à désinvestir les fonds UC pour sécuriser leur épargne sur le fonds euro (voir figure 1.13).

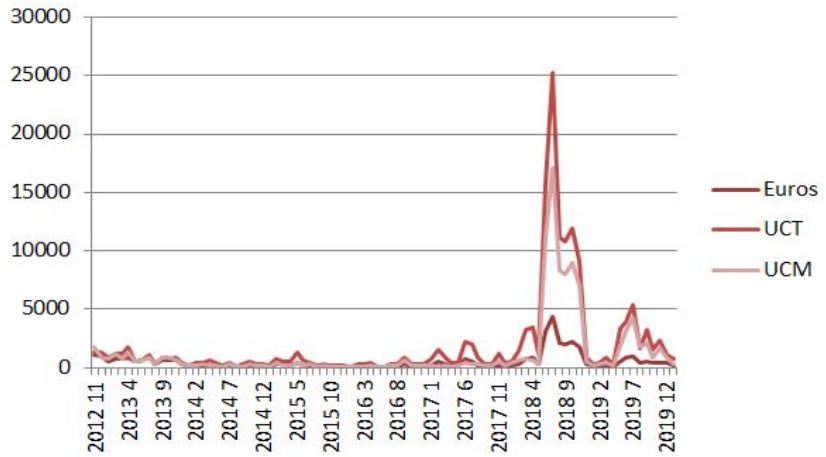

FIGURE 1.13 – Nombre d'arbitrages actés par type de fonds entre Novembre 2012 et Janvier 2020

Dans la suite, il s'agira de modéliser les mouvements d'arbitrage sur ce portefeuille et de dégager les profils des individus ayant un attrait particulier pour les arbitrages. Mais avant, il faut regarder quels sont les éléments qui influencent les arbitrages chez les assurés.

2. Les arbitrages : Etude des éléments impactants

Avant de s'attaquer à l'étude des éléments qui font varier les mouvements d'arbitrage, il est nécessaire de s'attaquer au traitement des données. Le but est d'obtenir une base de données fiable et complète pour tous les modèles par la suite.

2.1 Les données utilisées

2.1.1 Présentation de la base de données

La base de données finale est obtenue en regroupant les informations de 4 bases différentes :

1. **Une base contrat :**

   Elle comprend toutes les informations relatives aux mouvements d'arbitrage réalisés sur le contrat d'épargne du client. Pour capter au mieux le comportement des clients, la base a été réduite uniquement aux arbitrages volontaires expressément demandés par les clients. Les arbitrages automatiques liés à la gestion pilotée ont été supprimés. Un mouvement d'arbitrage est enregistré dans le système avec autant de lignes que de supports désinvestis et de supports investis. Chacune des lignes qui compose un arbitrage comprend les données suivantes :

   - ☐ Le numéro du contrat
   - ☐ Le produit associé : GPE ou GPV
   - ☐ La date de souscription du contrat
   - ☐ La date à laquelle l'arbitrage a été effectué
   - ☐ Le support sur lequel l'arbitrage a lieu
   - ☐ Le type d'arbitrage, c'est-à-dire s'il s'agit d'un arbitrage sortant du support ou d'un arbitrage entrant sur le support
   - ☐ Le montant de l'arbitrage : Il est signé négativement si l'arbitrage est sortant, et positivement si l'arbitrage est entrant
   - ☐ Le taux de frais prélevés pendant l'arbitrage
   - ☐ Le montant de l'encours sur le fonds euros
   - ☐ Le taux minimum annoncé (TMA) sur le fonds euros à la date de l'arbitrage : Il s'agit du taux utilisé pour calculer le montant des intérêts sur le fonds euros avant que le taux de PB soit connu
   - ☐ Le montant total de l'encours sur les fonds UC tiers
   - ☐ Le montant total de l'encours sur les fonds UC maison
   - ☐ Le montant total de l'encours sur le contrat (fonds euros et UC confondus)

     Les variables ci-dessous ont été calculées à partir des données déjà disponibles dans la base afin de la compléter :

   - ☐ Le mois de référence : Il correspond au mois et à l'année durant lesquels l'arbitrage a été effectué

- ☐ L'ancienneté du contrat : Elle est calculée à partir de la date de souscription et de la date d'arbitrage
- ☐ La variable Ind-Campagne : Elle vaut 1 pendant les périodes de business transformation et 0 sinon
- ☐ Les taux d'UC maison et d'UC tiers constituant l'épargne du client

Ci-dessous un exemple simplifié d'enregistrement d'arbitrage dans le système :

Numéro de contrat	**Date d'arbitrage**	**Support**	**Type d'arbitrage**	**Montant**
12121997	03/08/2020	Fonds euros	Sortant	- 10 000
12121997	03/08/2020	Fonds UC tiers 1	Entrant	+ 2 000
12121997	03/08/2020	Fonds UC tiers 2	Entrant	+ 4 000
12121997	03/08/2020	Fonds UC Generali	Entrant	+ 3 920

Table 2.1 – Exemple : Enregistrement d'un arbitrage dans la base

2. **Une base client :**

Elle comprend toutes les informations personnelles du client. Il s'agit des variables ci-dessous :

- ☐ L'identifiant du client
- ☐ La civilité du client : Indique si le client est un homme, une femme ou une personne morale
- ☐ La date de naissance du client
- ☐ L'âge au moment de l'abitrage : Il est calculé en faisant la différence entre la date de l'arbitrage et la date de naissance du client
- ☐ Les variables concernant le lieu de résidence du client : le département, le code postal, le numéro de commune INSEE [1], la ville (indique la place de la commune par rapport à l'agglomération urbaine à laquelle elle appartient : centre-ville, zone rurale, banlieue,...), le pays
- ☐ Le nombre de changements d'adresse du client
- ☐ La catégorie socio-professionnelle du client
- ☐ Le nombre de personnes constituant le foyer du client
- ☐ Refus de publicité : Elle vaut 1 si le client refuse toute forme de publicité et 0 sinon

Toutes ces informations sont en principe mises à jour chaque mois. Pour des raisons de volumétrie et de disponibilité, les informations n'ont pu être extraites qu'à la maille annuelle. La base contient tout l'historique des valeurs de novembre 2012 à mai 2020.

1. Institut national de la statistique et des études économiques

3. **Une base détention :**

   Elle répertorie toutes les informations concernant les autres contrats détenus par le client ainsi que les garanties qu'ils couvrent. Elle comprend les variables suivantes :

   - ☐ La date de passage d'un mono-équipement (détention d'un seul contrat) à un multi-équipement
   - ☐ La famille du dernier sinistre clôturé
   - ☐ La famille du dernier contrat souscrit
   - ☐ Le nombre d'affaires nouvelles souscrites par le client sur les 12 derniers mois
   - ☐ Le nombre de contrats détenus au total
   - ☐ Le nombre de contrats détenus pour différentes garanties : IARD, Vie, ADP, Autre DAB / RC, autre prévoyance, auto, DAB, Epargne-retraite, GAV, MRC, MRH, PNO, Santé, Emprunteur,...
   - ☐ Le nombre d'enfants du client
   - ☐ Le nombre de familles dans lesquelles le client détient au moins un contrat
   - ☐ Le nombre de contrats sortis en épargne-retraite, ADP et DAB
   - ☐ Le nombre de véhicules auto particuliers
   - ☐ Le nombre de véhicules auto professionnels
   - ☐ Le sexe du client
   - ☐ La situation familiale du client

     A partir de ces informations, les variables ci-dessous ont été obtenues :

   - ☐ Ind-multi : Elle vaut 1 si le client possède un multi-équipement au moment de l'arbitrage et 0 sinon

     De même que pour la base client, les informations de la base détention sont extraites année par année, de 2012 à 2020.

4. **Une base avec les données macroéconomiques :**

   Les bases contrat, client et détention ont été complétées avec des données relatives à l'environnement économique. Il s'agit de :

   - ☐ $CAC40_i, i = 0, 1, 2$: Performance mensuelle du CAC 40 i mois avant que l'arbitrage ait été effectué
   - ☐ TME : Valeur du taux moyen d'emprunt d'Etat le mois où l'arbitrage a été effectué
   - ☐ $Eurostoxx_i, i = 0, 1, 2$: Performance mensuelle de l'Eurostoxx 50 i mois avant que l'arbitrage ait été effectué
   - ☐ Le taux du livret A, principal concurrent bancaire de l'assurance-vie.

2.1.2 Constitution de la base de données

Toutes les bases de données présentées précédemment ont été fusionnées en une seule base en utilisant le mois de référence, le numéro de contrat ou encore le numéro de client comme clés pour les jointures (voir figure 2.1).

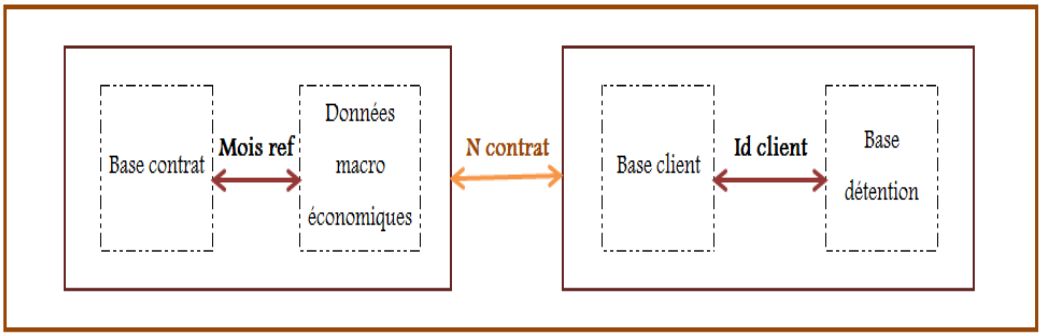

FIGURE 2.1 – Construction de la base de données

2.2 Traitement de la base

Après la fusion des bases de données, il a fallu procéder au traitement des données. Ce dernier s'est fait en deux étapes : d'abord le traitement des valeurs aberrantes, puis le traitement des valeurs manquantes.

2.2.1 Les valeurs aberrantes

Cette première étape correspond à la correction des erreurs de saisie constatées dans la base. Elle a relativement été rapide puisque la plupart des variables ne présentaient pas d'incohérences. Néanmoins, des difficultés sont apparues avec certaines variables telles que : le taux de frais d'arbitrages et le montant de l'encours. Ici, seul le cas des frais d'arbitrage sera présenté.

Lors de l'enregistrement d'un arbitrage dans la base, les frais d'arbitrage sont également renseignés. En reprenant l'exemple de la partie 2.1.1 de la page 32, l'enregistrement se présente comme suit :

N contrat	Date d'arbitrage	Support	Taux de frais	Frais	Montant arbitré
12121997	03/08/2020	Fonds euros	0.8 %	- 80	- 10 000
12121997	03/08/2020	Fonds UC tiers 1	0.0 %	+ 0.0	+ 2 000
12121997	03/08/2020	Fonds UC tiers 2	0.0 %	+ 0.0	+ 4 000
12121997	03/08/2020	Fonds UC Generali	0.0 %	+ 0.0	+ 3 920

Table 2.2 – Présentation simplifiée d'un mouvement d'arbitrage

Le taux de frais d'arbitrage est calculé sur les fonds de départ (ou sortants) en fonction des sommes arbitrées. Sur les fonds d'arrivée (ou entrants), il est automatiquement renseigné à 0%. Pour l'arbitrage du tableau 2.2.1 de la page 36, le taux appliqué est de $0,8\%$. Le montant en euros de frais d'arbitrage correspondant (-80 euros) est marqué dans la colonne *Frais* en négatif pour préciser qu'il s'agit bien d'un prélèvement.

Selon les conditions générales des contrats GPE et GPV, les frais appliqués lors d'un arbitrage ne peuvent excéder $0,8\%$ du montant total arbitré. En regardant les taux appliqués de plus près, les premières incohérences apparaissent :

Taux appliqué	Nombre
0.0 %	67 575
0.8 %	11 095
0.15 %	2 472
-0.01 %	114
100 %	29
-0.04 %	3
1.51 %	2

Table 2.3 – Aperçu des taux de frais dans la base

Certains arbitrages ont été réalisés avec des taux de frais dépassant le taux de frais maximum autorisé de $0,8\%$ (autre que 100%). Ils ont été supprimés car ils représentaient une proportion négligeable (0.4%) de la base de données.

Par ailleurs, une proportion assez importante d'arbitrages sont réalisés avec des taux de frais négatifs. Ils traduisent le fait que le client ait été rémunéré pour réaliser ces arbitrages. De nombreux arbitrages ont été réalisé avec un taux de 100%. Les taux de frais sont étudiés en parallèle avec le montant de frais imputés. Pour corriger ces mouvements, il a fallu procéder en plusieurs étapes :

— **Etape 1 :** Suppression des arbitrages actés avec un taux de frais d'arbitrage à 100%

Ces mouvements correspondent en réalité à des annulations d'arbitrages. Il est préférable de les retirer de la base car ils ne sont pas pertinents dans le cadre de cette étude. En plus des mouvements d'annulation qui ont été supprimés, les mouvements d'arbitrage initiaux ont également été effacés. Ceci permet de ne pas biaiser les résultats de l'étude par la suite.

— **Etape 2 :** Initialisation du taux d'arbitrage à 0% pour les arbitrages dont le montant de frais n'excède pas 1 euro.

Il est fréquent d'observer quelques centimes d'écart entre le montant arbitré réellement par le client et le montant arbitré renseigné dans la base. Cet écart est essentiellement dû aux erreurs d'arrondi des épargnes atteintes sur les différents fonds. De plus, en regardant de plus près, ces arbitrages partent du fonds euros vers les fonds UC tiers/ maison réalisés pendant les campagnes de business transformation. Il est donc parfaitement légitime de ramener le taux d'arbitrage appliqué à 0%. La figure 2.2 vous présente une estimation du nombre de mouvements d'arbitrage en fonction de la période.

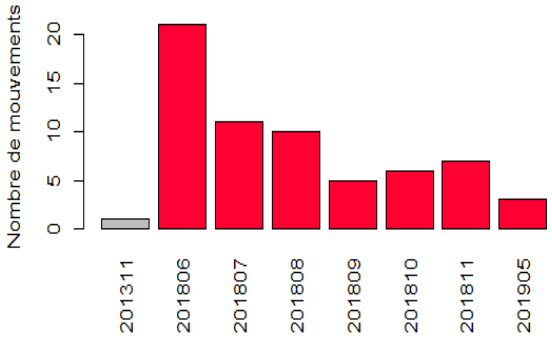

FIGURE 2.2 – Date des mouvements d'arbitrages à taux faibles

— **Etape 3 :** Correction des arbitrages dont les taux de frais sont connus

Parmi les arbitrages qui ont des taux d'arbitrage négatifs, certains ont été actés pendant les périodes de *business transformation*. Les frais d'arbitrage appliqués pendant ces périodes sont connus. Il a été aisé de les corriger.

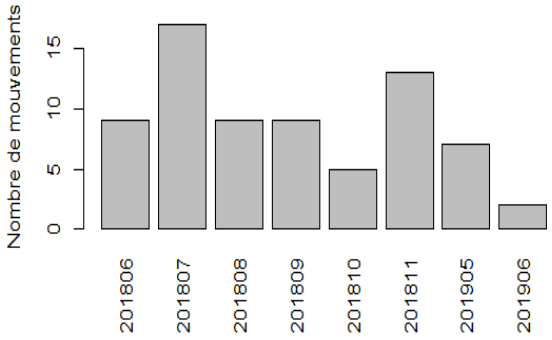

FIGURE 2.3 – Arbitrages négatifs réalisés uniquement pendant les campagnes d'arbitrages

Après ces traitements, les mouvements qui n'ont pas pu être corrigés ont tous été supprimés. La perte d'informations estimée pour la variable taux d'arbitrage à cette étape est inférieure à 1% .

2.2.2 Les valeurs manquantes

Les valeurs manquantes ou NA (*« not available »*) peuvent poser problème lors de l'utilisation de certains modèles. Elles sont traitées de différentes façons : elles sont soit exclues de la base d'apprentissage, soit remplacées par une valeur. Elles peuvent par exemple être remplacées par la moyenne des valeurs disponibles dans le jeu de données. L'exclusion des valeurs manquantes peut entraîner des pertes de précision, et même biaiser la calibration du modèle. L'imputation donne l'avantage de pouvoir travailler sur une base complète. Cependant, elle modifie la distribution des variables.

2.2.2.1 Les types de valeurs manquantes

Il existe plusieurs catégories de valeurs manquantes. Pour les classifier, il faut s'intéresser à leur cause. En fonction de si elles sont liées au hasard ou pas, Little et Rubin [8] les répartissent en 3 catégories :

— **Les valeurs manquantes totalement au hasard** ou **MCAR** (*missing completely at random*) : Une donnée est MCAR, c'est-à dire manquante de façon complètement aléatoire, si la probabilité d'absence est la même pour toutes les observations. Cette probabilité ne dépend donc que des paramètres extérieurs indépendants de cette variable. Par conséquent, la probabilité qu'une valeur soit manquante est une constante. Le cas des données MCAR est peu courant.

— **Les valeurs manquantes au hasard** ou **MAR** (*missing at random*) : Une valeur est manquante au hasard lorsque la probabilité d'absence des valeurs pour une observation ne dépend que des valeurs

observées. Cette probabilité dépend donc des autres variables.

— **Les valeurs non manquantes au hasard** ou **NMAR** (*not missing at random*) : Une valeur est non manquante au hasard lorsque la probabilité d'absence dépend de la variable ou des valeurs manquantes des autres variables.

Dans la pratique, il est difficile de savoir de quel type sont les valeurs manquantes. En ce qui concerne les deux premières catégories, c'est une connaissance métier qui permettra de déterminer l'appartenance. Par exemple, lors d'une enquête, les personnes aisées répondront moins que les autres à la question concernant leurs revenus. Pour les valeurs non manquantes au hasard, ce sera l'étude de la base qui pourra donner une première indication. Il faudra alors chercher des liens entre les variables qui présentent des valeurs manquantes.

2.2.2.2 Visualisation des données manquantes dans la base

La fonction *missmap* du package R *Amelia* permet d'obtenir une vision globale des valeurs manquantes. Elle permet ainsi d'observer d'éventuels liens entre les valeurs manquantes des différentes variables. Ci-dessous la matrice de données manquantes sur la base de données ainsi constituée (figure 2.4) :

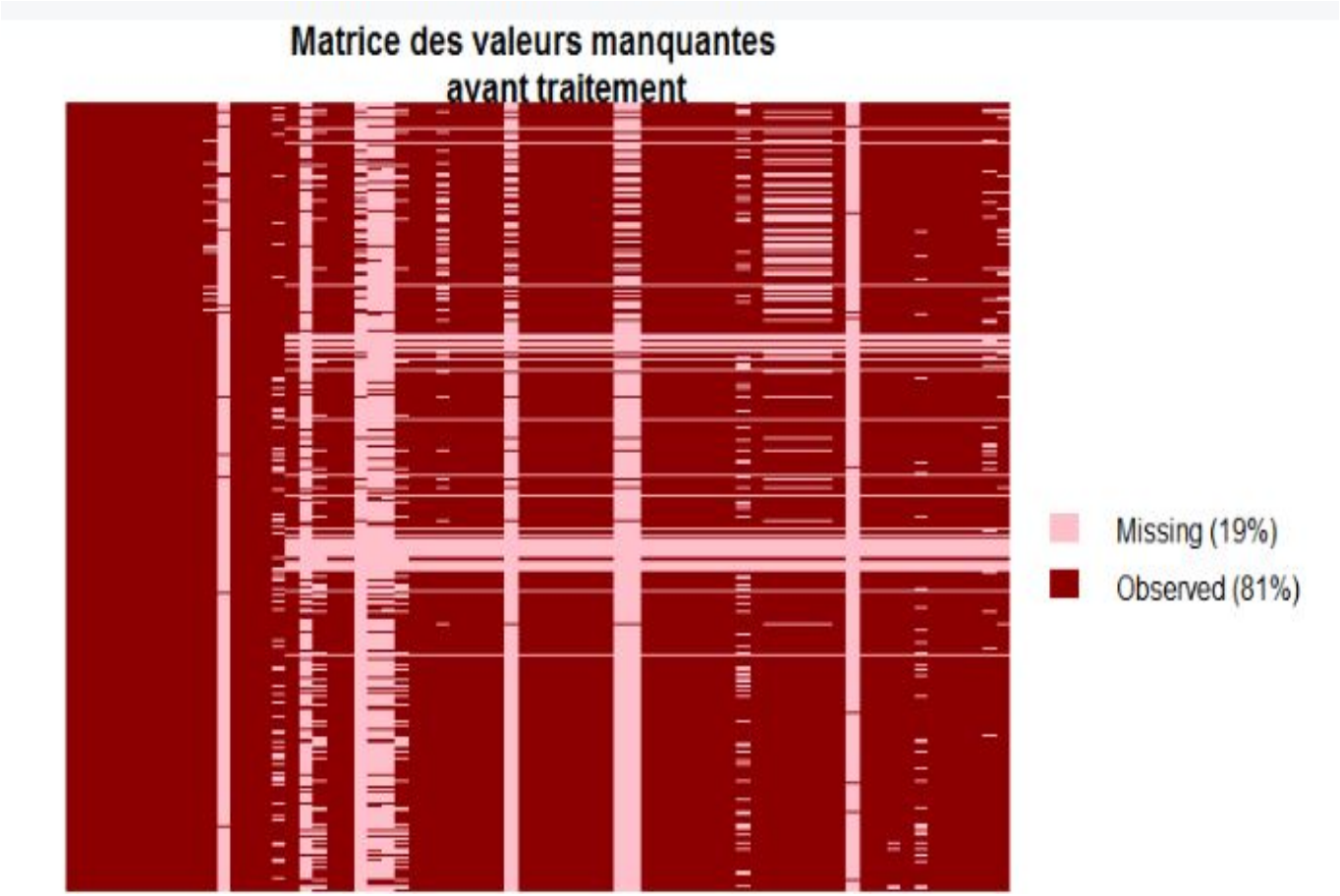

FIGURE 2.4 – Matrice des valeurs manquantes

Sur le graphique précédent [2], des liens entre les valeurs manquantes des différentes variables sont apparents. Ceci est essentiellement dû à la procédure utilisée pour constituer la base. Lorsque l'identifiant client est absent, systématiquement toutes les informations suivantes le sont aussi. Il y a ainsi plusieurs bandes de valeurs manquantes sur cette matrice.

2.2.2.3 Les méthodes d'imputation

La base comporte plusieurs types de variables : des variables qualitatives, des variables quantitatives et des variables de type date. En fonction du type de la variable et de la proportion de données manquantes, plusieurs méthodes d'imputation ont été utilisées :

1. **Imputation par la modalité la plus représentée**

2. Pour plus de lisibilité, les noms des variables ont été masqués

Cette méthode est principalement utilisée pour les variables qualitatives issues de la constitution de la base de données avec peu de valeurs manquantes. Les valeurs manquantes sont remplacées par la modalité la plus représentée si cela ne modifie pas trop les proportions des différentes modalités. Ci-dessous un tableau récapitulatif de quelques variables traitées avec cette méthode :

Variable	**Modalités**	**Répartition avant**	**Répartition après**
Nombre de contrats MRH	1 contrat	65.59 %	68.26 %
	2 contrats	32.59 %	30.06 %
	3 contrats	1.64 %	1.51 %
	4 contrats	0.16 %	0.14 %
Sexe	Homme	42.6 %	46.2 %
	Femme	57.4 %	53.8 %

TABLE 2.4 – Variables traitées avec la méthode d'imputation par la modalité la plus représentée

2. **Imputation par la moyenne**

   Cette méthode est utilisée pour des variables quantitatives avec peu de valeurs manquantes. Ici la valeur manquante est remplacée par la moyenne de la variable. C'est une méthode simple qui a l'avantage de ne pas modifier l'espérance. Elle a été utilisée avec précaution car elle biaise à la baisse la variance et peut fausser les résultats des tests statistiques par la suite.

3. **Création d'une nouvelle classe**

   — Pour les variables avec une grande proportion de valeurs manquantes :

     La création de cette nouvelle classe permet de capter un certain comportement chez les assurés. C'est le cas de la variable *Montant du revenu* avec un taux de valeurs manquantes de 14% pour laquelle la modalité "Non renseigné" a été crée. En effet, certains assurés riches préfèrent cacher le montant de leur revenu pensant qu'il impactera à la hausse le montant de leur prime.

   — Pour certaines variables de type date :

     Pour certaines variables, les valeurs manquantes ne sont pas forcément dérangeantes. Elles sont dues au renseignement de la variable dans la base. Il s'agit, par exemple de la variable *Date de passage multi-équipement*. Lorsqu'un assuré est toujours en mono-équipement, cette variable n'est pas renseignée.

4. **Imputation par KNN**

   Lorsque les méthodes précédentes ne sont pas applicables, la méthode des KNN (*k nearest neighbours*) est appliquée aussi bien pour les variables qualitatives que quantitatives. Pour les variables quantitatives, elle

consiste à affecter la valeur pondérée des k plus proches observations de la base ne comportant pas de valeurs manquantes. Concernant les variables qualitatives en revanche, la valeur manquante est remplacée par la classe la plus représentée chez les k plus proches voisins. Ainsi, pour chaque observation comportant des valeurs manquantes, il faut trouver les k individus les plus proches au sens d'une distance prédéfinie.

Dans le cadre des variables quantitatives, l'algorithme utilisé est le suivant :

Algorithme des k plus proches voisins (KNN)

(a) Choix d'un entier k, $1 \leq k \leq n$

(b) Calculer les distances $d(Y_{i^*}, Y_i)$, $i = 1, \ldots, n$

(c) Retenir les k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ pour lesquelles ces distances sont les plus petites

(d) Affecter aux valeurs manquantes la moyenne des valeurs des k voisins

$$y_{i,jmissing} = \frac{Y_{(i_1)} + \cdots + Y_{(i_k)}}{k}$$

Plusieurs paramètres sont à ajuster avant l'application de cet algorithme :

— **La distance**

Il existe plusieurs fonctions de calcul de distance : la distance euclidienne, la distance de Manhattan, la distance de Minkowski, la distance de Jaccard, la distance de Hamming ... La distance choisie dépend du type de données manipulées. Dans la mesure où les variables du jeu de données sont quantitatives, la distance euclidienne est un bon candidat. Cette dernière calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points. Pour deux points $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$, elle est définie comme suit :

$$Distance_{eucli}(x, y) = \sqrt{\sum_{i=1}^{n} (x_j - y_j)^2}$$

— **Le nombre de variables explicatives**

K-NN est un algorithme assez simple à appréhender. Principalement, grâce au fait qu'il n'a pas besoin de modèle pour pouvoir effectuer une prédiction. En revanche, il doit garder en mémoire l'ensemble des observations pour pouvoir effectuer sa prédiction. Ainsi, il faut faire attention à la taille du jeu d'entrée. Un nombre trop important de variables conduit à de mauvaises prédictions. Comme la base de données utilisée pour obtenir les prédictions comporte un grand nombre de variables, un tri a été effectué afin de sélectionner un nombre raisonnable de variables pour le modèle final. Le critère utilisé pour effectuer la réduction de dimension est le calcul du coefficient de corrélation. La définition de ce dernier est donnée ci-après :

Coefficient de corrélation linéaire :

Soit X et Y deux variables réelles de variances finies. Le coefficient de corrélation linéaire ρ_{XY} entre X et Y est défini comme :

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

où $Cov(X,Y)$, σ_X et σ_Y représentent respectivement la covariance de X et Y, l'écart-type de X et l'écart-type de Y.

Les variables retenues sont celles qui sont les plus corrélées avec la variable à prédire. Elles sont donc différentes d'un modèle à l'autre car la variable à prédire (qui est la variable à compléter) n'est pas la même.

Concernant le modèle ayant servi pour l'imputation des valeurs manquantes de la variable *Montant de l'épargne*, les variables explicatives retenues sont *Le montant de l'arbitrage* et *Le type de fonds*. En effet, les liens entre les autres variables explicatives et le montant de l'épargne sont négligeables. L'extrait de la matrice de corrélation ci-dessous le justifie :

Nom de la variable	**Valeur du coefficient de corrélation**
Montant de l'arbitrage	0.54
Type de fonds	0.23

TABLE 2.5 – Valeurs des coefficients de corrélation entre le montant de l'épargne et quelques variables explicatives

— **Le nombre k de voisins**

L'étape la plus délicate lors de l'application de l'algorithme KNN est le choix du nombre de voisins k. La valeur de k déterminée varie en fonction du jeu de données en entrée. Un nombre de voisins trop petit conduira à des problèmes de sous-apprentissage. Par contre, plus le nombre de voisins est élevé, plus la prédiction sera fiable. Toutefois, un nombre de voisins $k = n$ avec n la taille du jeu de données en entrée risquerait d'entraîner des problèmes de sur-apprentissage. Dans les deux cas, les conséquences sont de mauvaises performances pour le modèle prédictif.

Il existe plusieurs méthodes pour le choix du paramètre k. Ici, le nombre k de voisins est déterminé soit à l'aide du critère de minimisation du RMSE lorsque la variable à prédire est quantitative, soit du critère de maximisation du taux de précision des prédictions pour les variables qualitatives.

Le RMSE (*root mean square error*) ou racine carrée de l'erreur moyenne [3] est une mesure qui caractérise la précision d'une prédiction. Plus il est petit, plus la prédiction se rapproche de la réalité. Ainsi, la valeur de k retenue est celle qui conduit à la plus petite valeur de RMSE.

Le taux de précision, quant à lui, représente le pourcentage de prédictions exactes de notre modèle. Plus il est élevé, plus le modèle est performant. Par conséquent, la valeur de k retenue ici est celle qui conduit au

3. défini dans l'annexe 1 de la page 97

plus grand taux de précision.

Ci-dessous, les figures 2.5 (cas d'une prédiction qualitative) et 2.6 (cas d'une prédiction quantitative) sont deux exemples de choix de valeurs de k, l'un pour une variable qualitative et l'autre pour une variable quantitative :

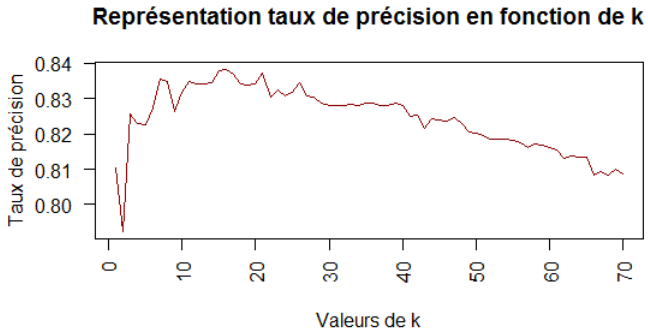

FIGURE 2.5 – Valeur de k = 16 retenue

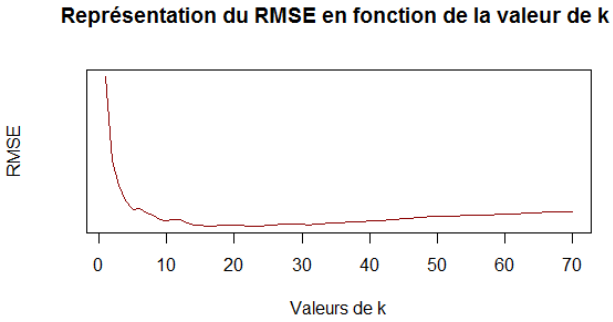

FIGURE 2.6 – Valeur de k = 22 retenue

2.3 Etude de l'impact des différentes variables

Maintenant que la base est complète, il s'agit de mieux la comprendre et regarder l'impact des différentes variables sur les flux arbitrés.

2.3.1 Réactivité des clients

La base de données constituée recense 22 178 mouvements d'arbitrages. Parmi ces arbitrages, 21 942 ont été actés par des clients détenteurs d'un contrat GPE et seulement 236 par des clients en possession d'un contrat GPV. En moyenne, les clients GPE réalisent 1,1 arbitrages (calculés avec un total de 20 035 contrats) contre une moyenne de 1,03 pour les clients GPV (pour un total de 230 contrats). Les assurés qui ont un contrat GPE sont donc plus sujets à faire des arbitrages que les autres. En regardant en terme de fréquence, la tendance se confirme :

Nombre d'arbitrages	GPE	GPV
11	1	0
10	1	0
7	1	0
6	3	0
5	2	0
4	15	0
3	157	0
2	1 500	6
1	18 355	224

Table 2.6 – Nombre de contrats par fréquence d'arbitrages pour GPE et GPV

D'après la table précédente, seulement 6 contrats GPV réalisent plus d'un arbitrage. Les clients GPE sont clairement plus réactifs que les clients GPV. Sur certains contrats GPE, le nombre d'arbitrages actés peut aller jusqu'à 11 pendant la période de l'étude. Même si la plupart des contrats se limitent à 1 ou 2 arbitrages.

2.3.2 Analyse des flux arbitrés

Dans cette partie, il s'agit de voir l'influence des différentes variables sur les flux arbitrés.

- **Le type de produit**

  En plus d'arbitrer plus souvent que les clients GPV, les clients GPE arbitrent des montants plus importants en moyenne. En effet, les clients GPV ont un encours moyen de 6300 euros contre un encours de 36 800 euros pour les clients GPE. Le type de produit a donc une influence sur le montant arbitré. La figure 2.7 ci-dessous montre le montant moyen arbitré pour chaque produit.

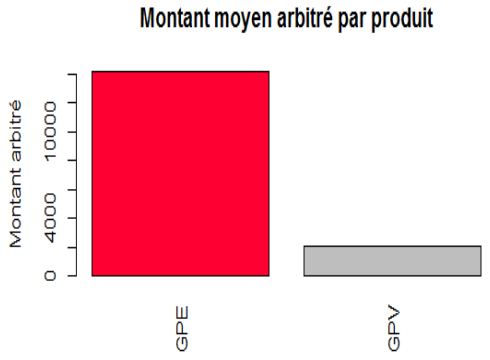

FIGURE 2.7 – Montant moyen arbitré pour les produits GPE et GPV

- **L'ancienneté du contrat**

  D'après la figure 2.8, les arbitrages sont effectués sur des contrats dont l'ancienneté varie entre 0 et 16 ans. 61 % des contrats ont entre 5 et 10 ans d'ancienneté. L'ancienneté est une variable importante dans la mesure où elle est directement liée à la fiscalité du contrat. Entre 0 et 4 ans, la fréquence d'arbitrage augmente sensiblement avec l'ancienneté. Après 4 ans d'ancienneté, il y a une hausse considérable au niveau du nombre d'arbitrages. Ceci peut s'expliquer par le changement au niveau de la fiscalité sur le contrat. Il y a ensuite un relâchement à partir de la 7ème année. Entre la 8ème et 9ème année d'ancienneté le nombre d'arbitrage décolle à nouveau. Au bout de 10 ans, les clients sont moins vifs et la fréquence d'arbitrage décroît jusqu'à la 16^{ieme} année.

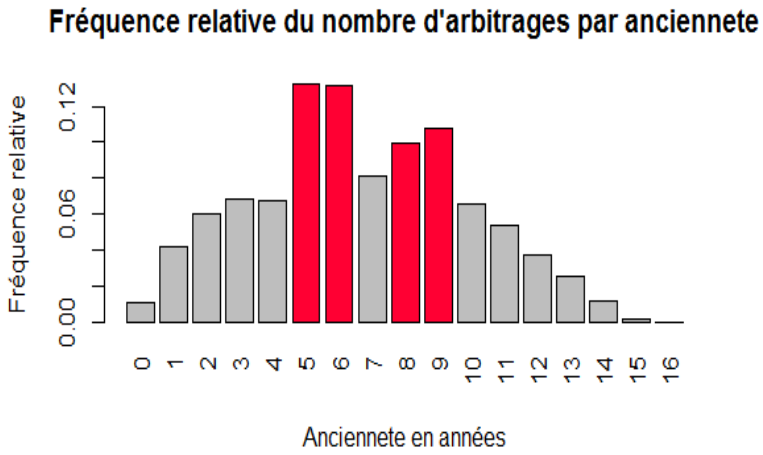

FIGURE 2.8 – Fréquence des arbitrages en fonction de l'ancienneté

Concernant le montant moyen arbitré, en se référant à la figure 2.9, il ressort globalement qu'avec l'ancienneté, les clients sont plus à l'aise. Ils arbitrent ainsi des montants moyens plus élevés. Au début, cette hausse est assez timide et, à partir de 6 ans, elle est beaucoup plus importante.

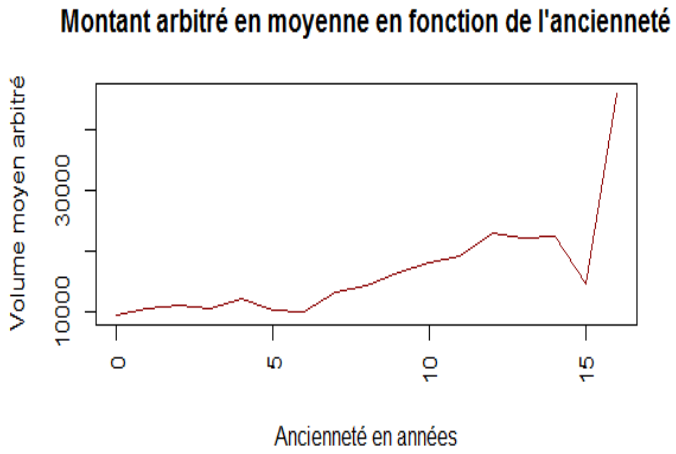

FIGURE 2.9 – Montant moyen arbitré en fonction de l'ancienneté

- **Le taux de frais d'arbitrage**

Après le traitement de la base, seuls trois taux appliqués sont présents : 0%, 0,15% et 0,8%.

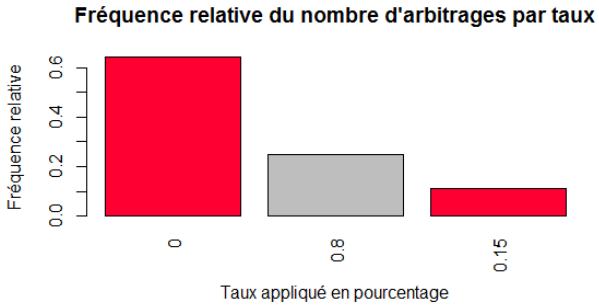

FIGURE 2.10 – Fréquence des arbitrages en fonction du taux d'arbitrage

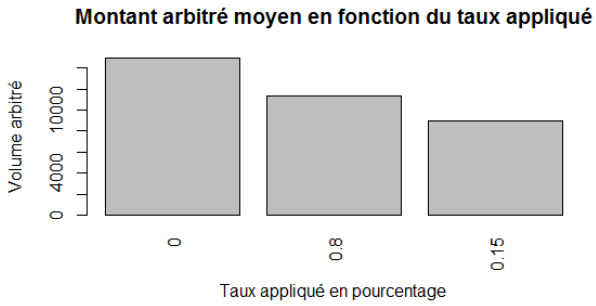

FIGURE 2.11 – Montant moyen arbitré en fonction du taux appliqué

Ces graphes 2.10 et 2.11 révèlent qu'environ 75% des arbitrages observés ont été effectués pendant les campagnes de *business transformation*. Les assurés n'étaient pas vraiment concernés par les arbitrages avant les campagnes. Ils ont profité des arbitrages gratuits proposés pendant les campagnes pour arbitrer en moyenne des montants plus grands.

- **La famille d'entrée**

  Elle représente la famille à laquelle appartient le premier contrat souscrit par le client. L'analyse de cette variable permet d'avoir une première idée sur l'impact de la multi-détention (voir figures 2.12 et 2.13).

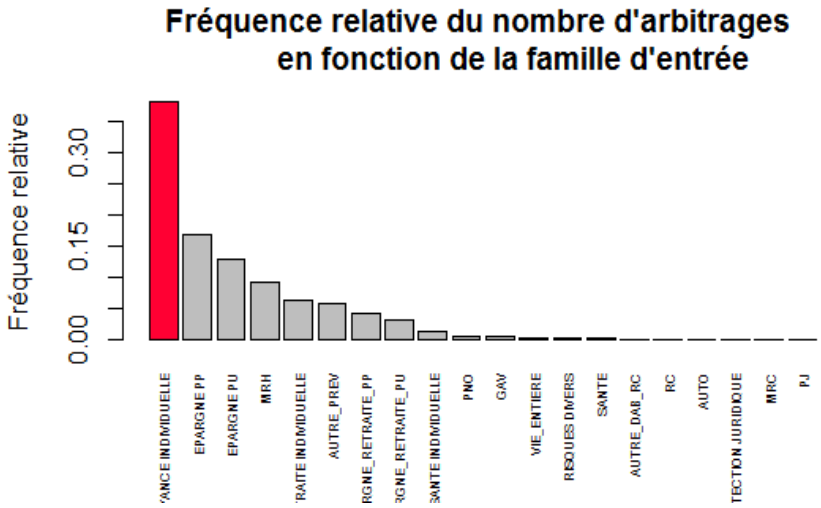

FIGURE 2.12 – Fréquence des arbitrages en fonction de la famille d'entrée

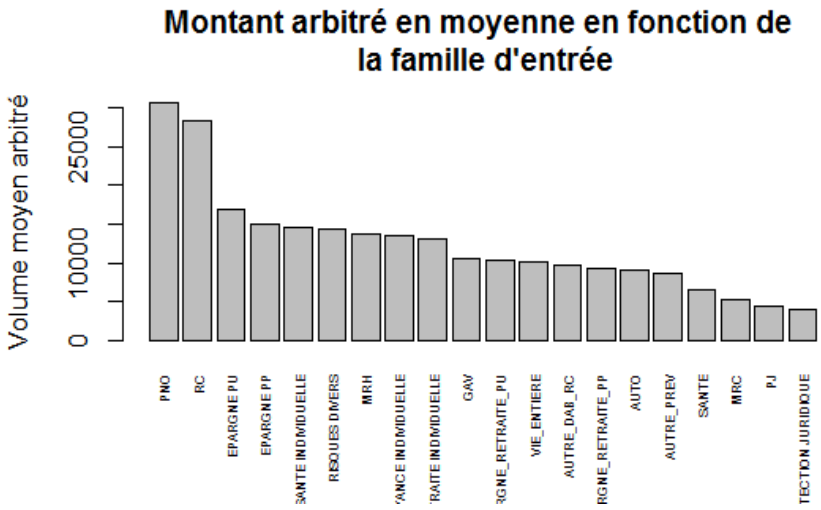

FIGURE 2.13 – Montant moyen arbitré en fonction de la famille d'entrée

Les arbitrages sont plus nombreux pour les clients qui ont des contrats de prévoyance individuelle en entrée. Ils sont suivis par les clients qui ont pris des contrats d'épargne en entrée (prime unique ou prime périodique). S'agissant des montants moyens arbitrés, il apparaît que les clients qui prennent en entrée des contrats RC [4] et PNO [5] arbitrent des montants plus importants en moyenne. Ces informations sont peu fiables compte tenu du nombre d'arbitrages observés pour ces familles. Les clients qui ont la famille épargne (PU et PP [6]) en entrée arbitrent des montants moyens un peu plus gros que les autres familles. Pour les plus petits montants moyens arbitrés, il faut regarder la famille protection juridique.

- **Le nombre de contrats total détenus**

  L'étude de cette variable est aussi réalisée dans l'optique de se faire une meilleure idée de l'impact de la multi-détention. La figure 2.14 montre que les clients qui effectuent des arbitrages ont à leur actif entre 1 et 25 contrats. Même si la grande majorité de ces clients détiennent entre 1 et 4 contrats au total. Il apparaît clairement que le nombre d'arbitrages actés diminue lorsque le nombre de contrats détenus augmente. Quand le nombre de contrats augmente considérablement, les assurés sont moins regardants sur leurs contrats d'épargne et effectuent ainsi moins d'arbitrages. Le nombre de contrats détenus n'a un impact que sur la fréquence des arbitrages.

4. Responsabilité civile
5. Assurance propriétaire non-occupant : Elle permet d'assurer un logement non-occupé par son propriétaire, qu'il soit loué ou vide, en cas de sinistre.
6. Prime unique et Prime périodique

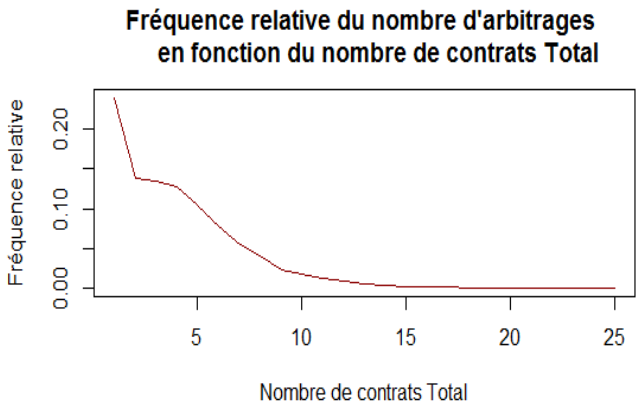

FIGURE 2.14 – Fréquence des arbitrages en fonction du nombre total de contrats

La figure 2.15 montre que les montants moyens arbitrés sont approximativement similaires, peu importe le nombre de contrats, et tournent autour de 15 000 euros. Le bruit observé pour un nombre de contrats élevé peut s'expliquer par un manque de données. A ce stade là, la multi-détention entraîne une baisse du nombre d'arbitrages.

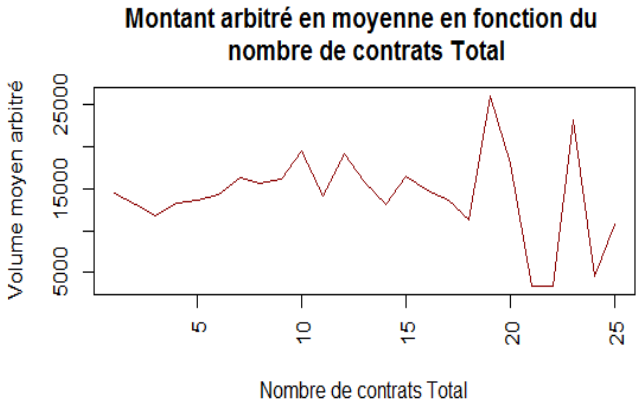

FIGURE 2.15 – Montant moyen arbitré en fonction du nombre total de contrats

- **Le sexe**

  Il ressort que les femmes arbitrent plus fréquemment que les hommes (voir figure 2.16). Cependant, le sexe n'a aucune influence sur le montant moyen arbitré puisque les deux sexes arbitrent le même montant moyen(voir figure 2.17).

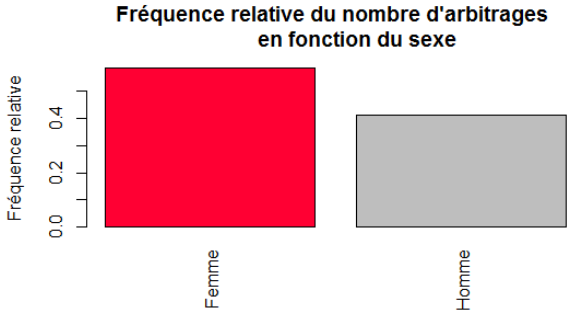

FIGURE 2.16 – Fréquence des arbitrages en fonction du sexe de l'assuré

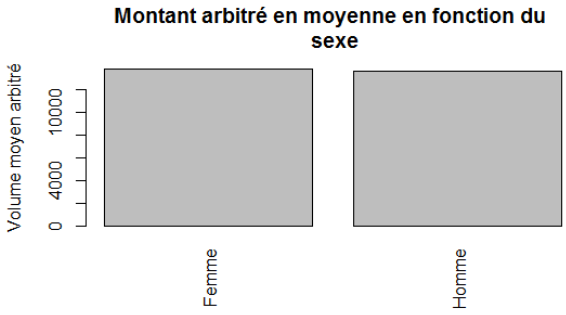

FIGURE 2.17 – Montant moyen arbitré en fonction du sexe de l'assuré

- **Recodage de certaines variables**

  En plus d'avoir une meilleure connaissance de la base, les représentations ci-dessus ont également permis de recoder des variables en regroupant certaines modalités. Les modalités dont les montants arbitrés, ou les fréquences d'arbitrages étaient très proches ont facilement été mises ensemble. C'est le cas, par exemple, de la variable *Ancienneté du dernier contrat souscrit* qui initialement était une variable quantitative dont les valeurs allaient de 0 à 15 ans. Après recodage, elle a été transformée en variable qualitative avec seulement 3 modalités : "0 an", "entre 1 et 5 ans" et "Plus de 5 ans écoulés". Les graphiques 2.18 et 2.19 ci-dessous justifient cette transformation :

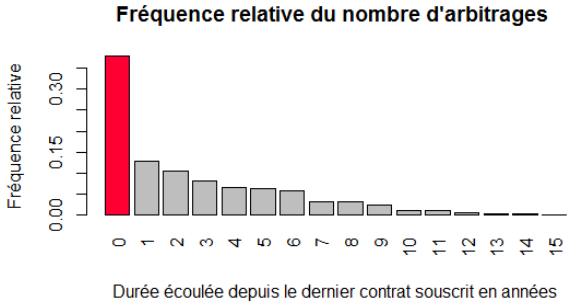

FIGURE 2.18 – Fréquence des arbitrages en fonction de la durée écoulée depuis le dernier contrat souscrit

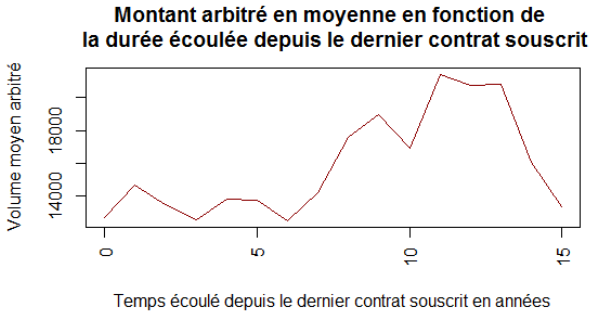

FIGURE 2.19 – Montant moyen arbitré en fonction de la durée écoulée depuis le dernier contrat souscrit

Les assurés qui ont souscrit un contrat dans l'année de l'arbitrage sont clairement plus vifs que les autres. Entre 0 et 5 ans d'ancienneté, le montant moyen arbitré est relativement stable, autour de 14 460 euros. A partir de 6 ans, il croît rapidement jusqu'à atteindre 21 435 euros.

2.4 Réduction des facteurs par ACP :

2.4.1 Principe

L'analyse en composantes principales (ACP) est une méthode mathématique d'analyse graphique de données. Elle consiste à remplacer une famille de variables par de nouvelles variables de variance maximale, non corrélées deux à deux et qui sont des combinaisons linéaires des variables d'origine. Ces nouvelles variables, appelées composantes principales, définissent des plans factoriels qui servent de base à une représentation graphique plane des variables initiales.

Cette méthode permet à la fois d'identifier les individus qui se ressemblent (notion de proximité) et de résumer les relations entre les variables.

Plus de détails sur cette méthode se trouvent dans [9].

2.4.2 Mise en oeuvre d'une ACP sur les données macro économiques

La base de données constituée comporte un trop grand nombre de variables explicatives. Un tel nombre de variables peut s'avérer problématique par la suite.

En regardant spécifiquement les variables concernant les données macroéconomiques, un phénomène intéressant se dégage. La matrice de corrélation (voir figure 2.20, page 53) révèle des liens importants entre nos variables. Pour éviter une redondance d'informations, une ACP a été réalisée.

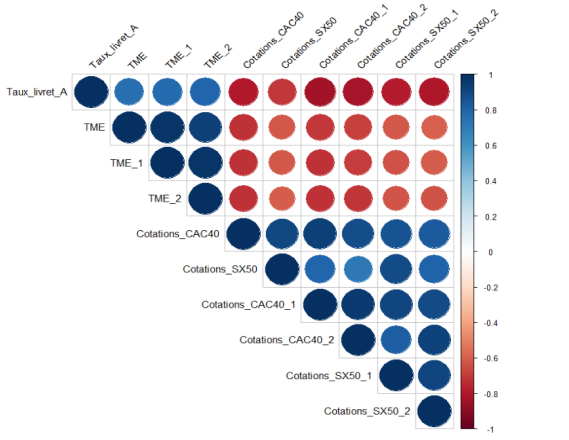

FIGURE 2.20 – Matrice de corrélation des variables macro économiques

L'ACP normée réalisée à l'aide du package R *ade 4*, conduit aux résultats suivant :

— **Le choix du nombre de facteurs :**

Il constitue une étape importante lors d'une ACP. Le but est d'obtenir un résumé suffisamment précis de l'information contenue dans le tableau initial tout en minimisant le nombre de facteurs retenus. Plusieurs critères de sélection existent, notamment le critère de Kaiser [7] ou encore le critère du coude. Ici, le critère du coude a été utilisé. Il s'appuie sur la règle ci-dessous :

Règle du coude :

Proposée par R. B. Cattell [8], elle consiste à étudier la courbe de décroissance des valeurs propres. L'idée est de détecter les « coudes », c'est-à-dire les cassures signalant un changement de structure. Les axes retenus sont ceux situés avant le coude.

Ainsi, en se référant à la figure 2.21 de la page 54, dans le cas des variables macro économiques utilisées, il sera sélectionné uniquement deux axes.

7. Il préconise de retenir uniquement les axes dont l'inertie est supérieure à l'inertie moyenne
8. 1905 - 1998, psychologue et professeur d'université anglo-américain

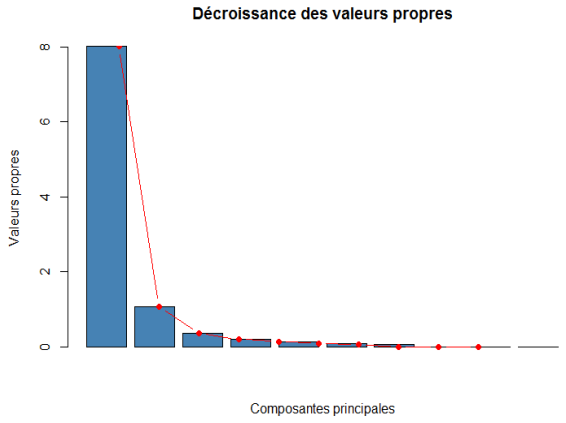

FIGURE 2.21 – Valeurs propres ACP - Données macro économiques

— **Interprétation des axes :**

Les deux premières composantes principales sont retenues pour la suite de l'étude. Ici, le premier axe factoriel extrait 80,07% de l'inertie totale. Le deuxième axe factoriel, quant à lui, récupère seulement 10.85% de l'inertie totale. Le premier plan factoriel représente à lui seul plus de 90% de l'inertie initiale. La projection du nuage de points initial sur le premier plan factoriel permet de conserver une grande partie de l'information de départ.

Le cercle de corrélation des différentes variables dans ce plan (figure 2.22 de la page 54) révèle une opposition, sur le premier axe, entre d'un côté les cotations du CAC 40 et de l'Eurostoxx 50 et leurs décalages dans le temps ; et, de l'autre côté, le taux du livret A et le TME et ses décalages. Il y a une corrélation négative entre les variables de ces deux groupes. Toutes les variables sont bien représentées sur cet axe. Sur le deuxième axe, en revanche, seuls les TME sont bien représentés.

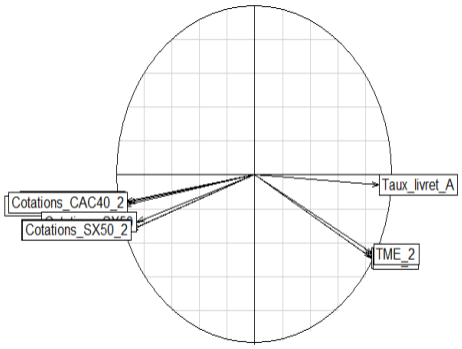

FIGURE 2.22 – Cercle de corrélation ACP - Données macro économiques

3. Modélisation des taux d'arbitrage à l'aide de différents modèles

3.1 Définition de la grandeur à modéliser

Les clients ont la possibilité de modifier la répartition de leur épargne entre les différents types de fonds : Euros, UCM et UCT. Seuls les arbitrages entre fonds euros sont impossibles (dans le cas de cette étude) car il n'y a qu'un seul fonds euros ouvert pour les produits GPE et GPV. Ceci laisse tout de même aux clients l'opportunité d'effectuer des arbitrages dans 8 sens différents [1].

Cependant, pour la suite de cette étude, les taux d'arbitrage seront modélisés dans 2 sens : du fonds euros vers les UC, et, des UC vers le fonds euros. Faute d'historique suffisant, les UCT et UCM ont été regroupées en une seule entité dite UC. Les arbitrages UC vers UC n'ont pas été modélisés car ils ne modifient pas la structure du portefeuille.

3.1.1 Une première approche

A partir de ce chapitre, les arbitrages ne seront plus évoqués en terme de volume ou de fréquence, mais plutôt en terme de taux. De plus, une distinction sera faite en fonction de la provenance et de la destination des flux arbitrés.

Cette perception des mouvements d'arbitrage par les taux d'arbitrage plutôt que par les montants d'arbitrage est préférable. En effet, elle permet de ne pas donner trop de primauté à des flux arbitrés importants, même si ceux-ci ne représentent qu'un pourcentage minime du montant de l'épargne du fonds d'origine.

En considérant un arbitrage qui a lieu d'un fonds d'origine A vers un fonds d'arrivée B, le taux d'arbitrage est défini comme ci-dessous à l'échelle du contrat :

$$\text{Taux d'arbitrage}_{A->B} = \frac{\text{Montant arbitré vers le fonds B}}{\text{Montant de l'épargne sur le fonds A}}$$

Le montant de l'épargne a été récupéré avant le mouvement d'arbitrage. Dans la mesure où il s'agit d'un désinvestissement du fonds d'origine, le montant arbitré ne peut excéder le montant de l'épargne (le client n'a pas la possibilité d'arbitrer un montant plus élevé que celui qui est disponible sur le fonds de départ). Ainsi, tous les taux d'arbitrage sont compris entre 0% et 100%. Un taux égal à 100% correspond à un désinvestissent total du fonds source. Un taux proche de 0% traduit en revanche un désinvestissement assez faible.

3.1.2 L'approche retenue

1. Euros vers UCT, Euros vers UCM, UCT vers Euros, UCT vers UCT, UCT vers UCM, UCM vers Euros, UCM vers UCT, UCM vers UCM

Lorsqu'ils veulent acter un arbitrage, les clients remplissent un bordereau dans lequel ils indiquent, pour chacun des fonds concernés par l'arbitrage, soit le montant, soit la proportion de l'épargne à investir ou à désinvestir au moment de l'arbitrage.

La première approche, certes naturelle et intuitive, n'a finalement pas été retenue pour deux principales raisons :

— Une nombre trop important d'arbitrages a été acté avec des proportions rondes. Une première représentation de la fonction de répartition des taux d'arbitrage observés révèle des sauts importants au niveau des taux ronds (par exemple 20 %, 30 %, ...90%, 100%). Les clients ont tendance à s'orienter plutôt vers ces taux là lorsqu'ils font une demande d'arbitrage. Ces sauts créent des paliers qui s'avèrent problématiques pour la suite de l'étude.

— Elle ne permet pas de faire la comparaison des taux d'arbitrage dans les deux sens retenus car les dénominateurs ne sont pas les mêmes.

Afin de contourner ces difficultés, les taux d'arbitrage sont finalement définis en fonction de l'épargne globale sur le contrat comme suit :

$$\text{Taux d'arbitrage}_{A->B} = \frac{\text{Montant arbitré vers le fonds B}}{\text{Montant de l'épargne globale sur le contrat}}$$

3.2 Échantillonnage de la base

Afin de définir le modèle le plus adapté à nos données, il est nécessaire de construire ce modèle et de le tester. Pour cela, chacune des deux bases de données (celle comprenant les arbitrages du fonds euros vers les UC et celle comprenant les arbitrages des UC vers le fonds euros) est divisée en deux parties :

— Une base d'apprentissage : constituée ici de 70% des données, celle-ci permet de construire le modèle et notamment d'ajuster au mieux tous les paramètres nécessaires ;

— Une base de test qui permet de définir l'erreur du modèle défini et de la comparer à l'erreur d'autres méthodes ou modèles. Elle compte pour 30% des données.

Ci-dessous un schéma représentant la manière dont sont utilisées les bases pour construire les modèles.

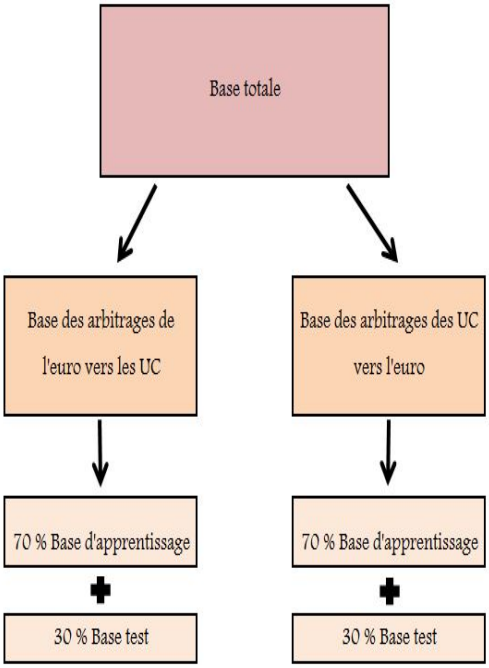

FIGURE 3.1 – Description du processus d'échantillonage adopté pour la modélisation des taux d'arbitrage

Il est à noter que les quatre échantillons obtenus (deux échantillons par type d'arbitrages) sont sauvegardés et utilisés pour tous les modèles implémentés afin d'avoir le même biais d'échantillonnage sur les résultats pour une comparaison optimale.

Pour la suite de cette étude, il est envisagé de modéliser les taux d'arbitrage à l'aide de plusieurs types de modèles d'abord simples, et ensuite plus complexes. Il s'agit notamment des modèles linéaires généralisés, des arbres CART [2] et des forêts aléatoires.

3.3 Le modèle linéaire généralisé (GLM)

Les modèles linéaires généralisés sont utilisés pour expliquer le comportement d'une variable continue Y par rapport à p prédicateurs ou variables explicatives $X_1, \ldots, X_p$.

3.3.1 Hypothèses

Ce type de modèles se caractérise par le choix de trois composantes :

2. CART : Classification and regression trees

1. Une composante aléatoire :

   Il s'agit du vecteur aléatoire à expliquer $Y = (Y_1, \ldots, Y_n)$. La densité de Y appartient à la famille exponentielle. Elle s'écrit de la façon suivante :

   $$f(y;\theta;\phi) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\}$$

   Avec θ le paramètre naturel de la famille, ϕ le paramètre de dispersion et a, b, c trois fonctions.

   Souvent $a(\phi)$ est remplacé par $\frac{\phi}{\omega_i}$ avec ω_i un poids attribué à la i^{eme} observation.

   Plusieurs lois usuelles appartiennent à la famille exponentielle telles que les distributions normale, exponentielle, gamma et bêta.

2. Une composante déterministe :

   Il s'agit d'une combinaison linéaire des vecteurs explicatifs $X_1, \ldots, X_p$.

3. Une fonction lien g :

   C'est une fonction réelle, déterministe et strictement monotone telle que :

   $$g_n(E(Y)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

   Où g_n est une fonction de $\mathbf{R}^n$ vers $\mathbf{R}^n$ définie par $g_n(x_1, \ldots, x_n) = (g(x_1), \ldots, g(x_n))$

De plus, si Y appartient à la famille exponentielle, alors :

$$\mu = E(Y) = b'(\theta)$$
$$Var(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$$

Où $V(\mu)$ est appelé fonction variance.

3.3.2 Estimation des paramètres

Une fois les différentes composantes sélectionnées, le but est d'estimer les paramètres du modèle. Ces derniers sont calculés à partir de la méthode du maximum de vraisemblance.

Soient $X_1, \ldots, X_p$ les p vecteurs explicatifs et g la fonction lien choisie. Alors :

$$\forall i \in [\![1\ ;\ n]\!], g(E(Y_i)) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

Or

$$E(Y_i) = b'(\theta_i)$$

Avec b' inversible et g bijective.
Il est donc possible d'exprimer θ_i en fonction de $\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$, $\forall i \in [\![1\ ;\ n]\!]$.

En effet :

$$\begin{aligned}\theta_i &= (b')^{-1}(E(Y_i)) \\ &= T(E(Y_i))\end{aligned}$$

Avec $E(Y_i) = g^{-1}(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi})$

$$= (Tog^{-1})(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi})$$

Dans la fonction de vraisemblance $(\theta_1, \ldots, \theta_n)$ sera remplacé par la valeur obtenue ci-dessus.

En notant les estimateurs obtenus par la méthode du maximum de vraisemblance des paramètres $(\beta_0, \ldots, \beta_p)$ et ϕ par $(\hat{\beta}_0, \ldots, \hat{\beta}_p)$ et $\hat{\phi}$ respectivement, la valeur ajustée par le modèle pour l'observation y_i sera :

$$\hat{y_i} = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots, \hat{\beta}_p X_{pi})$$

L'équation de log-vraisemblance s'écrit :

$$\begin{aligned}l(\theta(\beta), y, \phi) &= \sum_{i=1}^{n} \ln(f(y_i, \phi, \theta_i)) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{\omega_i}} + \sum_{i=1}^{n} c_i(y_i, \phi)\end{aligned}$$

Pour $j = 0, \ldots, p$, $\dfrac{\partial l(\theta(\beta), y, \phi)}{\partial \beta_j} = 0$

Or,

$$\frac{\partial \ln(f(y_i, \phi, \theta))}{\partial \beta_j} = (\frac{\partial \ln(f(y_i, \phi, \theta_i))}{\partial \theta_i} / \frac{\partial \mu_i}{\partial \theta_i}) \times \frac{\partial \mu_i}{\partial \beta_j}$$

Chacun des termes individuellement devient :

$$\begin{aligned}\frac{\partial \ln(f(y_i, \phi, \theta_i))}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{\frac{\phi}{\omega_i}} \\ &= \frac{y_i - \mu_i}{\frac{\phi}{\omega_i}}\end{aligned}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

$$\begin{aligned}\frac{\partial \mu_i}{\partial \beta_j} &= \frac{\partial \mu_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_j} \\ &= x_{ij} \frac{\partial \mu_i}{\partial \eta_i}\end{aligned}$$

Le résultat devient alors :

$$\frac{\partial \ln(f(y_i, \phi, \theta))}{\partial \beta_j} = \frac{(y_i - \mu_i) x_{ij} \frac{\partial \mu_i}{\partial \eta_i}}{\frac{\phi}{\omega_i} b''(\theta_i)}$$

Finalement, les équations de vraisemblance s'écrivent :

$$\sum_{i=1}^{n} \omega_i (y_i - \mu_i) \frac{x_{ij}}{b''(\theta_i) g'(\mu_i)} = 0, \forall j = 0, \ldots, p$$

En général, il n'existe pas de solutions explicites pour ces équations. Il faut avoir recours à des méthodes plutôt numériques, telles que la méthode de Newton-Raphson [3], ou alors, la méthode du score de Fisher [4] (qui utilise la matrice d'informations de Fisher).

3.3.3 Validation des paramètres

Une fois les paramètres estimés pour un modèle donné, il faut s'intéresser à la significativité des estimateurs obtenus ainsi qu'à la significativité du modèle. Pour cela, deux approches peuvent être considérées : le test du rapport de vraisemblance ou encore le test de Wald [5].

— **Le test du rapport de vraisemblance**

La statistique du ratio de vraisemblance permet de vérifier la pertinence du modèle. L'hypothèse nulle est « Les coefficients associés aux variables explicatives sont tous nuls ». Elle se traduit par :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

L'hypothèse alternative est « Il existe un des coefficients associés aux variables explicatives qui est non nul ». Elle se traduit comme ci-dessous :

$$H_1 : \exists i \in [\![1 \ ; \ k]\!] \text{ tel que } \beta_i \neq 0$$

3. Des mathématiciens Isaac Newton (1643 - 1727) et Joseph Raphson 1648 - 1715. Il s'agit d'un algorithme efficace pour trouver numériquement une approximation précise d'un zéro (ou racine) d'une fonction réelle d'une variable réelle. Plus d'informations sur cet algorithme en annexe.

4. Ronald Aylmer Fisher (1890 - 1962), biologiste et statisticien britannique

5. Abraham Wald (1902 - 1950), mathématicien américain qui contribua à la théorie de la décision statistique, à la géométrie à l'économétrie et fonda le domaine de l'analyse séquentielle statistique

En posant L_1 et L_0 les vraisemblances des modèles sans et avec contraintes respectivement, le rapport de vraisemblance est défini comme :

$$\lambda = \frac{L_1}{L_0}$$

La statistique du test du rapport de vraisemblance est :

$$\Lambda = 2\ln(\lambda) = 2(l(\beta_1) - l(\beta_0))$$

Avec $l(\beta_1)$ et $l(\beta_0)$ les log vraisemblances des modèles sans et avec contraintes respectivement.

Sous l'hypothèse H_0, cette statistique suit une loi du khi deux à k degrés de liberté. L'hypothèse H_0 est rejetée lorsque la p-value associée est plus petite qu'un seuil fixé arbitrairement.

— **Le test de Wald**

Le test de Wald permet de vérifier la significativité d'un coefficient du modèle. Pour un coefficient β_j, l'hypothèse H_0 est « Le coefficient β_j est nul » :

$$H_0 : \beta_j = 0$$

L'hypothèse alternative est « Le coefficient β_j est non nul » :

$$H_1 : \beta_j \neq 0$$

La statistique du test de Wald est définie par :

$$\Lambda = \frac{\hat{\beta_j}^2}{Var(\hat{\beta_j})}$$

Cette statistique suit une loi du chi deux à 1 degré de liberté. Comme précédemment, la valeur de la p-value et celle du seuil fixé permettront de décider si l'hypothèse H_0 est rejetée ou pas.

3.3.4 Résidus du modèle

Il existe plusieurs types de résidus pour un modèle donné.

— **Les résidus lignes :**

Ils sont définis comme $r_i = y_i - \mu_i$. Les résidus ligne empiriques sont calculés comme suit : $\hat{r_i} = y_i - \hat{\mu_i}$

— **Les résidus de Pearson** [6] :

6. Karl Pearson 1857 - 1936, mathématicien britannique. C'est un des fondateurs de la statistique moderne appliquée à la biomédecine

Ils sont donnés par la formule : $r_i^P = \dfrac{\sqrt{\omega}_i(y_i - \mu_i)}{\sqrt{V(\mu_i)}}$ et par la formule : $\hat{r_i}^P = \dfrac{\sqrt{\omega}_i(y_i - \hat{\mu_i})}{\sqrt{V(\hat{\mu_i})}}$ pour les résidus de Pearson empirique.

Il est important de préciser que la somme des carrés des résidus suite une loi khi deux à $n - p - 1$ degrés de liberté.

— **Les résidus de déviance :**

En considérant que chaque observation y_i contribue à la hauteur d'une quantité d_i à la déviance du modèle D ($D = \sum_i^n d_i$), les résidus de déviance sont définis comme :

$$r_i^D = signe(y_i - \mu_i) \times \sqrt{d_i}$$

A noter que comme dans le cas des résidus de Pearson, la somme des carrés des résidus de déviance est asymptotiquement, un khi deux à $n - p - 1$ degrés de liberté.

La représentation graphique des résidus permet aussi d'avoir une indication sur la qualité du modèle.

3.3.5 Programmation sous R

Sous le logiciel R, l'implémentation d'un modèle GLM se fait à l'aide de la fonction *glm*. Elle prend en entrée les paramètres suivants :

— Y : la variable à prédire ;
— X : l'ensemble des variables explicatives ;
— *data* : la base de données à exploiter ;
— *family* : le paramètre qui permet d'indiquer la distribution de la famille exponentielle sélectionnée.

Les sorties R d'un modèle glm sont :

— Les valeurs des coefficients estimés ;
— Les résidus ;
— Les p-values associées aux différents tests de significativité des coefficients estimés.

Dans les deux sections suivantes, le lecteur trouvera des informations sur les arbres CART et les forêts aléatoires. Au besoin, plus de détails sur ces algorithmes se trouvent dans la bibliographie[7].

3.4 Les arbres CART

L'acronyme CART signifie ***C**lassification **A**nd **R**egression **T**rees*. Il désigne une méthode statistique, introduite par Breiman et al. [5] qui construit des prédicteurs par arbre aussi bien en régression qu'en classification. Les arbres CART sont la brique de base de deux algorithmes très connus : les forêts aléatoires (*random forest*) et les *gradient boosting trees*.

3.4.1 Principes

Les arbres CART ont pour but la classification ou le clustering des individus d'une population en différents groupes homogènes. Ils se définissent par :

— des branches et un tronc : Ils sont caractérisés par les règles de segmentation de la population ;

— d'un racine : Elle est composée de toute la population étudiée ;

— des feuilles : Elles contiennent les sous-populations finales et fournissent ainsi une estimation de la quantité d'intérêt.

3.4.2 Construction des arbres CART

Pour construire un arbre CART, il est nécessaire d'introduire deux règles de contrôle pour la division de la population :

1. Ne segmenter à chaque étape que sur **une** variable ;
2. La réponse à la question posée pour segmenter la population est binaire.

Plus simplement, les étapes à suivre pour la construction d'un arbre CART sont les suivantes :

— Étape 1 : Prendre une covariable et la partitionner. Si elle est quantitative, les partitionnements possibles se situent entre deux valeurs successives observées pour la variable. Sinon, la variable est qualitative, et les partitionnements possibles sont toutes les combinaisons possibles entre les modalités ;

— Étape 2 : Tester ces partitionnements et calculer un critère d'homogénéité basé sur la variable à expliquer ;

— Étape 3 : Choisir le partitionnement qui conduit à la plus grande homogénéité globale dans les deux sous espaces créés ;

— Étape 4 : Répéter les étapes 1 à 3 pour les variables explicatives à disposition. Cela permet d'obtenir plusieurs mesures d'homogénéité maximales, chacune correspondant à une covariable ;

— Étape 5 : Sélectionner la variable et son partitionnement qui maximise l'homogénéité.

L'arbre construit en suivant cette démarche correspond à l'arbre maximal. Il faut chercher à l'élaguer, c'est-à-dire, supprimer certaines branches afin de limiter le sur-apprentissage.

En réalité, bâtir un arbre CART nécessite deux étapes. Il y a tout d'abord une première phase de construction de l'arbre maximal qui permettra de définir la famille à l'intérieur de laquelle le meilleur modèle sera sélectionné. Puis, il y a une seconde phase dite d'élagage, qui construit une suite de sous arbres optimaux élagués de l'arbre maximal. Chacune de ces deux étapes est expliquée plus en détail ci-dessous.

3.4.2.1 Construction de l'arbre maximal

Dans cette partie, il s'agit de revenir de façon plus formelle sur la règle de découpe des branches lors de la construction de l'arbre maximal. Afin de mieux se fixer les idées, le lecteur peut se restreindre à des variables explicatives continues. Le cas des variables qualitatives est toutefois mentionné explicitement dans le texte chaque fois que cela est utile.

Dans le cas où, il y a p variables explicatives, l'espace d'entrée est supposé être $\mathbf{R}^p$. En partant de la racine de l'arbre (donc tout l'espace $\mathbf{R}^p$ ici), qui contient toutes les observations de l'échantillon d'apprentissage $\mathcal{L}_n$, la première étape de CART consiste à découper au mieux cette racine en deux nœuds fils. La coupure (ou découpe, ou même *split*) est un élément de la forme :

$$\{X_j \leq d\} \cup \{X_j > d\}$$

Où $j = 1, \ldots, p$ et $d \in \mathbf{R}^p$.

Découper suivant $\{X_j \leq d\} \cup \{X_j > d\}$ signifie que toutes les observations avec une valeur de la j^{ime} variable plus petite que d vont dans le nœuds fils de gauche, et toutes celles avec une valeur plus grande que d vont dans le nœuds fils de droite. La méthode sélectionne alors la meilleure découpe, c'est-à-dire le couple (j, d) qui minimise une certaine fonction de coût :

— En régression, le but est de minimiser la variance intra-groupes résultant de la découpe d'un nœud t en 2 nœud fils t_L et t_R. La variance d'un nœuds t étant définie par :

$$V(t) = \frac{1}{t} \sum_{i:x_i \in t} (y_i - \bar{y}_t)^2$$

Avec $\bar{y}_t$ la moyenne des y_i des observations présentes dans le nœud t. Cela conduit donc à minimiser :

$$\frac{1}{n} \sum_{(x_i,y_i) \in t_L} (y_i - \bar{y}_{t_L})^2 + \frac{1}{n} \sum_{(x_i,y_i) \in t_R} (y_i - \bar{y}_{t_R})^2 = \frac{t_L}{n} V(t_L) + \frac{t_R}{n} V(t_R)$$

— En classification (où l'ensemble des classes est $\{1, \ldots, L\}$), l'impureté des nœuds fils se caractérise le plus souvent par le biais de l'indice de Gini [7]. L'indice de Gini d'un nœud t est défini par :

7. Du statisticien italien Corrado Gini 1884 - 1965. La mesure d'impureté de Gini est l'une des méthodes utilisées dans les algorithmes d'arbre de décision pour décider de la division optimale à partir d'un nœud, et des divisions ultérieures.

$$\Phi(t) = \sum_{c=1}^{L} \hat{p}_t^c (1 - \hat{p}_t^c)$$

Avec $\hat{p}_t^c$ la proportion d'observations de classe c dans le nœud t.

Cela conduit donc à maximiser :

$$\Phi(t) - (\frac{t_L}{t}\Phi(t_L) + \frac{t_R}{t}\Phi(t_R))$$

En régression, l'objectif est de trouver des découpes qui tendent à minimiser la variance des nœuds obtenus. En classification, il s'agit plutôt de diminuer la fonction de pureté de Gini, et donc à augmenter l'homogénéité des nœuds obtenus. Un nœud est parfaitement homogène s'il ne contient que des observations de la même classe.

Dans le cas d'une variable explicative catégorielle X_j, rien de ce qui précède ne change sauf que dans ce cas, une coupure est simplement un élément de la forme :

$$\{X_j \in d\} \cup \{X_j \in \bar{d}\}$$

Où d et $\bar{d}$ sont non vides et constituent une partition de l'ensemble des modalités de la variable X_j.

Une fois la racine de l'arbre découpée, il faut s'intéresser à chacun des nœuds fils. Et, suivant le même procédé, l'idée est de trouver la meilleure façon de les découper en deux nouveaux nœuds, et ainsi de suite. Les arbres sont ainsi développés, jusqu'à atteindre une condition d'arrêt. Une règle d'arrêt classique consiste à ne pas découper des nœuds qui contiennent moins d'un certain nombre d'observations.

3.4.2.2 Élagage

La deuxième étape de l'algorithme CART s'appelle l'élagage. Elle consiste à chercher le meilleur sous-arbre élagué de l'arbre maximal (noté T_{max}) au sens de l'erreur de généralisation. L'idée est que l'arbre maximal possède une très grande variance et un biais faible. A contrario, un arbre constitué uniquement de la racine (qui engendre alors un prédicteur constant) a une très petite variance mais un biais élevé. L'élagage est une procédure de sélection de modèles, où les modèles sont les sous-arbres élagués de l'arbre maximal, soit tous les sous-arbres binaires de T_{max} ayant la même racine que T_{max}. Cette procédure minimise un critère de pénalité proportionnel au nombre de feuilles de l'arbre.

Tous les sous-arbres binaires de T_{max} contenant la racine sont des modèles admissibles. Entre T_{max}, le modèle de complexité maximale, qui conduit au surajustement aux données de l'échantillon d'apprentissage et l'arbre restreint à la racine qui est fortement biaisé, il s'agit de trouver l'arbre optimal parmi les admissibles. En nombre fini, il suffirait donc au moins en principe, de construire la suite de tous les meilleurs arbres à k feuilles pour $k = 1, \ldots, |T_{max}|$ où $|T|$ désigne le nombre de feuilles de l'arbre T, et de les comparer par exemple sur un échantillon test. Mais le nombre de modèles admissibles est exponentiel d'où une complexité algorithmique explosive. Fort heureusement, une énumération implicite et efficace suffit pour atteindre un résultat optimal. Le moyen consiste simplement dans l'algorithme d'élagage qui assure l'extraction d'une

suite de sous-arbres emboîtés (c'est-à-dire élagués les uns des autres) $T_1, \ldots, T_K$ tous élagués de T_{max} où T_k minimise un critère des moindres carrés pénalisé en régression. Cette suite est obtenue de manière itérative en coupant des branches à chaque étape, ce qui ramène la complexité à un niveau très raisonnable. Les lignes qui suivent se rapportent uniquement au cas de la régression, la situation étant identique en classification.

La clé est de pénaliser l'erreur d'ajustement d'un sous-arbre T élagué de T_{max} :

$$\bar{err}(T) = \frac{1}{n} \sum_{\{t \text{ feuille de } T\}} \sum_{(x_i, y_i) \in t} (y_i - \bar{y}_t)^2$$

par une fonction linéaire du nombre de feuilles $|T|$ conduisant au critère des moindres carrés pénalisés :

$$crit_\alpha(T) = \bar{err}(T) + \alpha |T|$$

Ainsi, $\bar{err}(T)$ qui mesure l'ajustement du modèle T aux données, décroît avec le nombre de feuilles alors que $|T|$ qui quantifie la complexité du modèle T, croît avec le nombre de feuilles. Le paramètre α règle la pénalité : plus α est grand, plus les modèles complexes (c'est-à-dire comptant beaucoup de feuilles) sont pénalisés.

L'algorithme d'élagage est donné dans la table 3.1, où pour tout nœud interne t d'un arbre T, la branche de T issue du nœud t (contenant tous les descendants du nœud t) est notée T_t, et, l'erreur correspondante est donnée par $\bar{err}(T) = n^{-1} \sum_{x_i \in t} (y_i - \bar{y}_t)^2$.

Entrée	Arbre maximal T_{max}
Initialisation	$\alpha_1 = 0$, $T_1 = T_{\alpha_1} = argmin_T$ élagué de $_{T_{max}} \overline{err}(T)$
	Initialiser $T = T_1$ et $k = 1$
Itération	Tant que $\lvert T \rvert > 1$,
	Calculer
	$\alpha_{k+1} = \min_{\{t \text{ nœud interne de } T\}} \dfrac{\overline{err}(t) - \overline{err}(T_t)}{\lvert T_t \rvert - 1}$
	Élaguer toutes les branches T_t de T telles que $\overline{err}(T_t) + \alpha_{k+1}\lvert T_t \rvert = \overline{err}(t) + \alpha_{k+1}$
	Prendre T_{k+1} le sous-arbre élagué ainsi obtenu
	Boucler sur $T = T_{k+1}$ et $k = k + 1$
Sortie	Arbres $T_1 \succ \cdots \succ T_K = \{t_1\}$,
	Paramètres ($0 = \alpha_1; \ldots; \alpha_K$)

TABLE 3.1 – Algorithme d'élagage de CART

Le résultat principal du livre de Breiman et al. [5] établit que la suite de paramètres $(0 = \alpha_1; \ldots; \alpha_K)$ est strictement croissante, associée à la suite $T_1 \succ \cdots \succ T_K = \{t_1\}$ constituée de modèles emboîtés au sens de

l'élagage et que, pour tout $1 \leq k \leq K$

$$\begin{aligned} \forall \alpha \in [\alpha_k; \alpha_{K+1}[\quad T_k &= \underset{\{T \text{ sous-arbre de } T_{max}\}}{\text{argmin}} \quad crit_\alpha(T) \\ &= \underset{\{T \text{ sous-arbre de } T_{max}\}}{\text{argmin}} \quad crit_{\alpha_k}(T) \end{aligned}$$

En posant ici $\alpha_{K+1} = \infty$.

Autrement dit, la suite $T_1, \ldots, T_K$ contient toute l'information statistique utile puisque pour tout $0 \leq \alpha$, le sous-arbre minimisant $crit_\alpha$ est un sous-arbre de la suite produite par l'algorithme d'élagage.

3.4.3 Programmation sous R

Sous le logiciel R, l'implémentation d'un arbre CART peut se faire à l'aide de la fonction *rpart*. Elle prend les paramètres suivant en entrée :

- $Y \quad X$: La formule à utiliser avec Y la variable à expliquer et X l'ensemble des variables explicatives sélectionnées ;
- *data* : La base de données à exploiter ;
- *control* : Il contient les paramètres :
  - *minsplit* : Le nombre minimum d'observations au niveau d'un nœud pour faire une coupure ;
  - *minbucket* : Le nombre d'observations minimum dans une feuille ;
  - *cp* : Le paramètre de complexité ;
  - *maxdepth* : La taille maximale de l'arbre ;
- *params* : La fonction d'impureté.

La sortie de cette fonction permet de récupérer le modèle CART avec plusieurs éléments : l'architecture de l'arbre, les différentes conditions de coupure utilisées, le vecteur des variables trié en fonction de leur pouvoir discriminant, etc.

3.5 Les forêts aléatoires

Les forêts aléatoires sont introduites par L. Breiman[2] en 2001. Comme le terme "forêt" l'indique, ces dernières sont composées d'un ensemble d'arbres décisionnels. Elles agrègent les arbres de décision de façon à augmenter leur robustesse tout en conservant leur coté intuitif.

3.5.1 Principe

Les forêts aléatoires font partie des méthodes ensemblistes [8]. Elles partent du principe que pour obtenir un résultat de meilleure qualité, il est nécessaire de recueillir au préalable les résultats de plusieurs modèles. Dans ce cadre là, plutôt que de se fier à un seul estimateur complexe, le but est de construire plusieurs estimateurs de moins bonne qualité individuelle. Chaque estimateur a ainsi une vision partielle du problème et tente de le résoudre au mieux avec les données dont il dispose. Puis, ces modèles aux estimations divergentes (car construits sur des données différentes) sont réunis pour fournir une vision globale. C'est l'agrégation de tous ces modèles qui rend les forêts aléatoires extrêmement efficaces. En régression, cette agrégation se fait par la moyenne des prédictions.

Afin d'amoindrir la corrélation entre les modèles et ainsi de diminuer la variance du modèle agrégé, l'algorithme des forêts aléatoires effectue un double échantillonnage. En effet, il y a non seulement un échantillonage des observations, mais aussi un échantillonage des variables. Les individus et les variables sont tous les deux tirés aléatoirement. L'algorithme dans lequel seul le tirage des individus est randomisé est connu sous le nom de *tree bagging*. Les forêts aléatoires ajoutent au *tree bagging* une randomisation sur les variables explicatives, appelée *feature sampling*.

3.5.1.1 Le bagging

La méthode du *bagging* a été introduite par L. Breiman[1] en 1996. Le mot *bagging* est la contraction des mots ***B**oostrap* et ***Agg**regat**ing***.

Soient un échantillon d'apprentissage $\mathcal{L}_n$ et une méthode de prédiction dite règle de base qui construit sur $\mathcal{L}_n$ un prédicteur $\hat{h}(., \mathcal{L}_n)$. Le bagging se fait en 3 étapes. Il consiste :

1. A tirer de façon indépendante plusieurs échantillons bootstrap $(\mathcal{L}_n^{\Theta_1}, \ldots, \mathcal{L}_n^{\Theta_q})$;
2. Ensuite, à appliquer la règle de base sur chacun d'eux afin d'obtenir une collection de prédicteurs $(\hat{h}(., \mathcal{L}_n^{\Theta_1}), \ldots, \hat{h}(., \mathcal{L}_n^{\Theta_q}))$;
3. Et enfin, à agréger ces prédicteurs de base.

L'idée du *bagging* est de modifier les prédictions en appliquant la règle de base sur différents échantillons bootstrap. Ceci, dans le but d'obtenir une collection variée de prédicteurs. L'étape d'agrégation permet finalement d'obtenir un prédicteur performant. La figure 3.2 ci-dessous illustre le *bagging* avec pour règle de base un arbre CART :

8. Une méthode ensembliste combine les décisions individuelles de plusieurs modèles de façon à améliorer les résultats et la robustesse

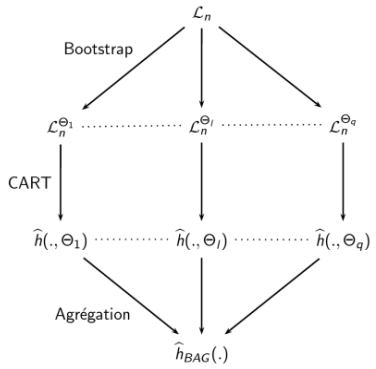

FIGURE 3.2 – Le *bagging* avec pour méthode de base les arbres CART

3.5.1.2 Les forêts aléatoires *Random Inputs*

Les forêts aléatoires -RI peuvent être vues comme une variante du *bagging*. La différence intervient dans la construction des arbres individuels. En effet, les étapes de bootstrap et d'agrégation restent identiques pour les deux algorithmes. Le tirage à chaque nœud des m variables se fait sans remise, et uniformément parmi toutes les variables. Le nombre m ($m \leq p$) est fixé au début de la construction de la forêt. Il est donc identique pour tous les arbres et pour tous les nœuds d'un même arbre. Mais, naturellement, les m variables impliquées dans deux nœuds distincts sont en général différentes. C'est un paramètre très important de la méthode. Une forêt construite avec $m = p$ revient à faire du *bagging* d'arbres CART non élagués, alors qu'une forêt construite avec $m = 1$ est très différente du *bagging*.

En pratique, les forêts aléatoires -RI améliorent les performances du *bagging* si le paramètre m est bien choisi. Le fait de rajouter un aléa supplémentaire pour construire les arbres rend ces derniers encore plus différents les uns des autres sans pour autant dégrader de façon significative leurs performances individuelles. Le prédicteur agrégé est alors meilleur.

La figure 3.3 ci-dessous fournit le schéma récapitulatif de l'algorithme des forêts aléatoires -RI, où Θ désigne le tirage bootstrap et Θ' désigne le tirage aléatoire des variables.

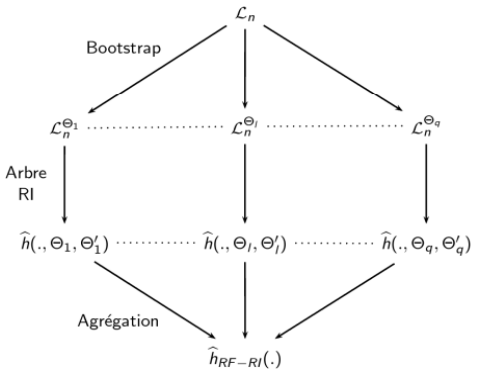

FIGURE 3.3 – L'algorithme random forests - RI

3.5.2 L'erreur OOB

L'algorithme des forêts aléatoires -RI calcule une estimation de son erreur de généralisation. Il s'agit de l'erreur *Out-Of-Bag* (OOB) où "Out-Of-Bag" signifie "en dehors du bootstrap".

Soit une observation (X_i, Y_i) de l'échantillon d'apprentissage $\mathcal{L}_n$. Cette observation est supposée out-of-bag, n'appartenant pas à l'ensemble des arbres construits sur les échantillons bootstrap. A partir des prédictions agrégées de ces arbres, il est possible d'obtenir la prediction $\hat{Y}_i$ de Y_i. Après avoir fait cette opération pour toutes les données de $\mathcal{L}_n$, l'erreur OOB est définie comme :

— L'erreur quadratique moyenne en régression donnée par la formule :

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

— La proportion d'observations mal classées en classification, définie par :

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{Y}_i \neq Y_i}$$

Cette estimation s'impose les mêmes contraintes que celles des estimateurs de l'erreur de généralisation du modèle CART (échantillon test, validation croisée), au sens où les données prédites sont des données qui n'ont pas été rencontrées au préalable par le prédicteur utilisé. Un avantage de l'erreur OOB par rapport aux estimateurs CART est qu'elle ne nécessite pas de découper l'échantillon d'apprentissage, puisque, le découpage est en quelque sorte inclu dans la génération des différents échantillons bootstrap. Cependant, il faut noter que pour chaque observation, il ne s'agit pas du même ensemble d'arbres agrégés.

3.5.3 L'importance des variables

L'avantage principal des arbres CART est leur interprétabilité. Une fois l'arbre construit, les variables discriminantes sont aisément identifiées. Il s'agit de celles qui interviennent effectivement dans les découpes des nœuds de l'arbre. Leur niveau d'importance est aussi facile à détecter : plus une variable est proche de la racine de l'arbre, plus elle est importante. Cette lisibilité se perd dans les méthodes ensemblistes comme les forêts aléatoires. Une forêt étant une agrégation de plusieurs arbres, le prédicteur obtenu n'est plus structuré.

Léo Breiman[2] introduit un autre indice d'importance des variables pour pallier à ce problème. Comme pour l'erreur OOB, le calcul de cet indice d'importance utilise les échantillons bootstrap.

Ci-après le calcul de l'indice de l'importance de la variable X_j pour $j = 1, \ldots, p$. Soit un échantillon bootstrap $\mathcal{L}_n^{\Theta_l}$ et l'échantillon OOB_l associé (ie. l'ensemble des observations qui n'apparaissent pas dans $\mathcal{L}_n^{\Theta_l}$). L'erreur commise sur OOB_l par l'arbre construit sur $\mathcal{L}_n^{\Theta_l}$ est notée err_{OOB_l}. En permutant alors aléatoirement les valeurs de la j^{ime} variable dans l'échantillon OOB_l, l'échantillon $\widetilde{OOB_l}^j$ est obtenu. Il s'agit ensuite de calculer l'erreur $err_{\widetilde{OOB_l}^j}$ sur l'échantillon $\widetilde{OOB_l}^j$. Pour finir, il faut répéter ces opérations pour tous les échantillons bootstrap. L'importance de la variable X_j, notée $VI(X_j)$, est définie par la différence entre l'erreur moyenne d'un arbre sur l'échantillon OOB perturbé et celle sur l'échantillon OOB :

$$VI(X_j) = \frac{1}{q} \sum_{i=1}^{q} (err_{\widetilde{OOB_l}^j} - err_{OOB_l})$$

Avec q le nombre d'échantillons bootstrap contenant la variable.

Ainsi, plus les permutations aléatoires de la j^{ieme} variable engendrent une forte augmentation de l'erreur, plus la variable est importante. A l'inverse, si les permutations n'ont quasiment aucun effet sur l'erreur (voire même la diminuent, ce qui fait que VI peut être légèrement négative), la variable est considérée comme très peu importante.

3.5.4 Programmation sous R

Le package R *RandomForest* peut être utilisé. Il nécessite un certain nombre de paramètres dont les principaux sont :

- *ntree* : Le nombre d'arbre agrégés. Il est initialisé à 500 par défaut ;
- *max depth* : La profondeur de l'arbre ;
- *mtry* : Le nombre $q < p$ de prédicteurs sélectionnés pour la construction de chaque modèle (p étant le nombre total de variables explicatives). Par défaut, il est égal à $\sqrt{p}$ en classification et $p/3$ en régression.

3.6 Les résultats

Dans cette section, il est question de présenter les résultats obtenus avec les différents modèles. Mais avant, la démarche utilisée pour choisir les paramètres des modèles sera présentée. Au total, en comptant 3 modèles pour chaque sens d'arbitrage, il a fallu établir 6 modèles. Pour des raison de simplicité, la démarche sera présentée uniquement pour les modèles des arbitrages du fonds euro vers les fonds UC. Concernant les arbitrages des fonds UC vers le fonds euro, toutes les informations relatives aux étapes permettant de calibrer les modèles se trouvent en annexe.

Avant de calibrer les différents modèles, il est important de regarder les données à modéliser. La figure 3.4 ci-dessous représente les taux moyens observés entre les années 2012 et 2020 pour les arbitrages du fonds euro vers les fonds UC :

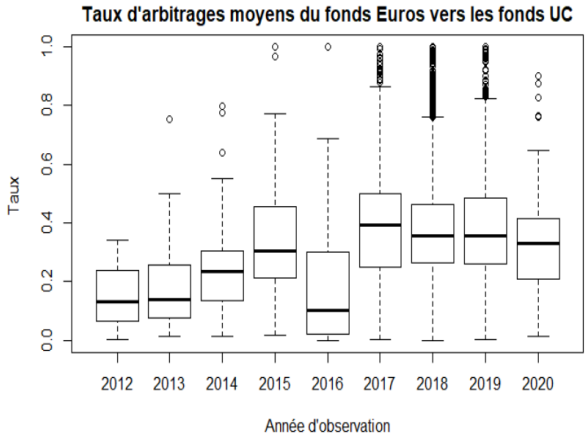

FIGURE 3.4 – Les taux moyens d'arbitrage observés pour les arbitrages du fonds euro vers les fonds UC

Entre les années 2012 et 2015, il y a une augmentation des taux moyens d'arbitrages. Et au cours de l'année 2016, cette moyenne de taux d'arbitrage observée sur le portefeuille baisse considérablement. Entre les années 2015 et 2016, elle perd environ la moitié de sa valeur et passe de 35% en 2015 à 17% en 2016. Les assurés deviennent un peu plus frileux pendant l'année 2016 et préfèrent arbitrer en moyenne des proportions moins importantes du fonds euro vers les fonds UC. Ceci peut s'expliquer par la baisse observée sur les marchés financiers au cours de l'année 2016. Il faut cependant noter que cette tendance à la baisse est très vite balayée pendant les années suivantes. A partir de l'année 2017, le taux moyen des arbitrages remonte. Entre les années 2017 et 2020, il gardera une valeur entre 32% et 40%. Pendant les années 2018 et 2019, durant lesquelles les clients ont bénéficié des campagnes de *business transformation*, il y a un nombre assez important de valeurs aberrantes. Ceci est sûrement dû au grand nombre d'arbitrages actés pendant cette période.

3.6.1 Choix des paramètres pour le modèle GLM : Arbitrages du fonds euro vers les fonds UC

Le modèle GLM comporte plusieurs paramètres. Pour obtenir le modèle le plus performant possible, il est important de choisir ces paramètres avec soin.

3.6.1.1 Choix de la distribution des taux d'arbitrage

Le calibrage du modèle GLM commence ici par le choix de la distribution des taux d'arbitrage. Plusieurs lois usuelles appartenant à la famille exponentielle ont été testées. Il s'agit des distributions exponentielle, gamma et normale. Le choix de la distribution adéquate s'est fait au regard de différents éléments tels que :

— **La comparaison des fonctions de répartition :**

Ici, il est nécessaire de comparer la fonction de répartition empirique et les fonctions de répartition des lois usuelles sélectionnées. La loi qui aura le plus de chances d'être retenue pour le modèle final sera celle qui se rapprochera au maximum de la loi empirique. Les graphes 3.5, 3.6 et 3.7 ci-dessous permettent de se faire une idée :

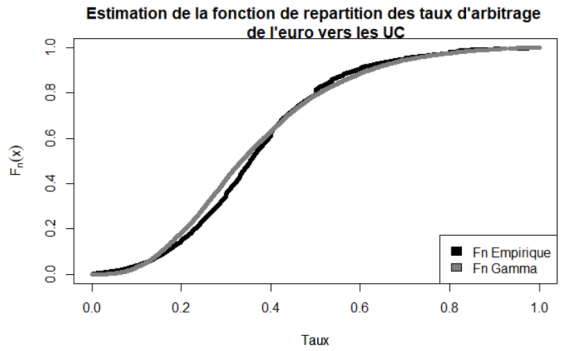

FIGURE 3.5 – Comparaison avec la fonction de répartition de la loi gamma

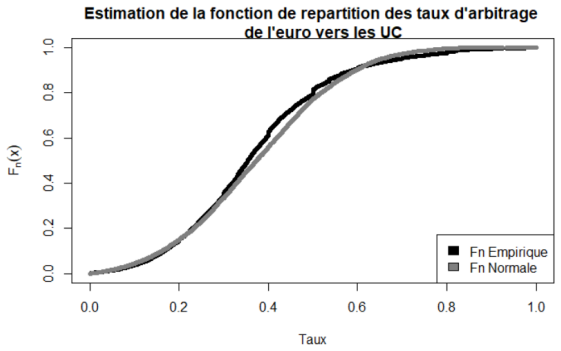

FIGURE 3.6 – Comparaison avec la fonction de répartition de la loi normale

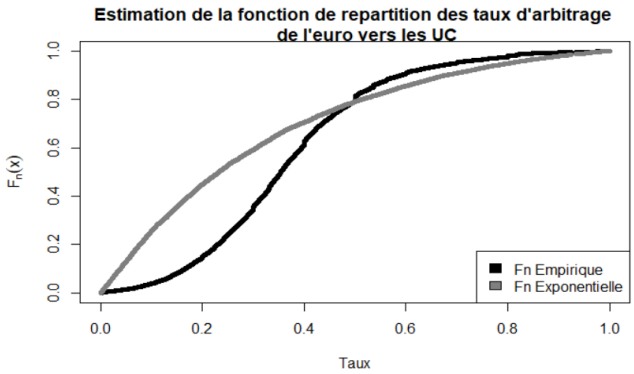

FIGURE 3.7 – Comparaison avec la fonction de répartition de la loi exponentielle

La figure 3.7 ci-dessus permet d'éliminer rapidement l'hypothèse de la distribution exponentielle. En effet, la fonction de répartition de la loi exponentielle est beaucoup trop éloignée de celle de la distribution empirique. A ce stade, le choix se porte plus vers les distributions gamma et normale, plus proches de la distribution observée. L'étape suivante permettra de faire un choix plus aiguisé entre ces deux distributions.

— **La comparaison des densités :**

Après la comparaison des fonctions de répartition, l'étude se poursuit avec la comparaison des densités. Il est question de comparer la densité empirique des taux d'arbitrage observés avec les densités des lois usuelles appartenant à la famille exponentielle. Et comme précédemment observé, la loi dont la densité se rapprochera le plus de la densité empirique sera la plus susceptible d'être choisie. La figure 3.8 ci-dessous montre les différentes densités :

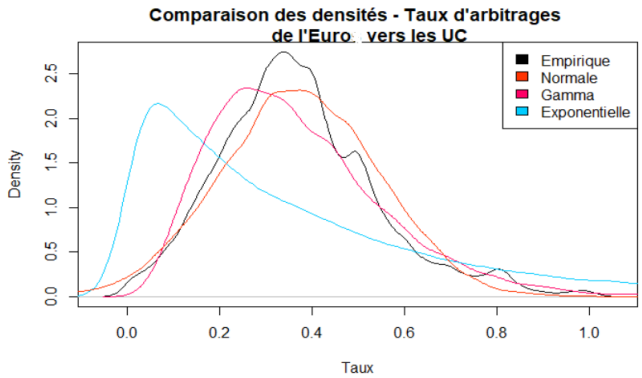

FIGURE 3.8 – Comparaison de la densité empirique des taux d'arbitrage et des densités des lois usuelles de la famille exponentielle

La figure 3.8 confirme qu'il faut écarter l'hypothèse d'une distribution exponentielle. Là aussi, la densité de la loi exponentielle est beaucoup trop éloignée de la densité empirique. Cette dernière présente une grosse bosse haute et des bosses moins prononcées en descendant. La densité de la loi exponentielle, plutôt symétrique, s'ajuste moins bien au niveau de la plus grosse bosse. La densité de la loi gamma semble être la meilleure option. Elle capte bien l'énorme bosse et se rapproche assez bien de la densité empirique des taux d'arbitrage en redescendant.

— **Représentation des QQ-plot [9] :**

9. Le QQ-plot (ou diagramme quantile-quantile) est un graphique qui permet d'évaluer la similarité de deux ensembles de données. Ici, le but est de confronter les données observées (les taux d'arbitrage) et des données théoriques (issues des lois usuelles de la famille exponentielle, à savoir les lois gamma, exponentielle et normale). Pour cela, il faut comparer la position de certains quantiles dans les données observées avec leur position dans la population théorique.

Un point de coordonnées (x,y) sur le graphique représente un quantile de la distribution observée (axe y) contre le même quantile de la distribution théorique (axe x). Lorsque la distribution théorique correspond à la distribution observée, les points sont parfaitement alignés sur la droite théorique. Les graphiques 3.9, 3.10 et 3.11 ci-dessous correspondent aux QQ-plots pour les lois gamma, exponentielle et normale :

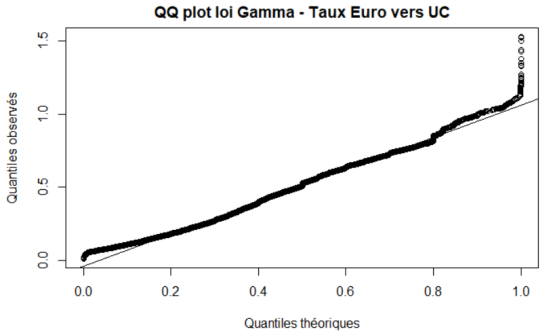

FIGURE 3.9 – QQ plot - Loi Gamma

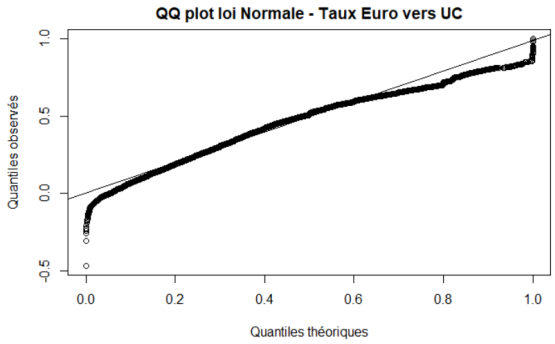

FIGURE 3.10 – QQ plot - Loi normale

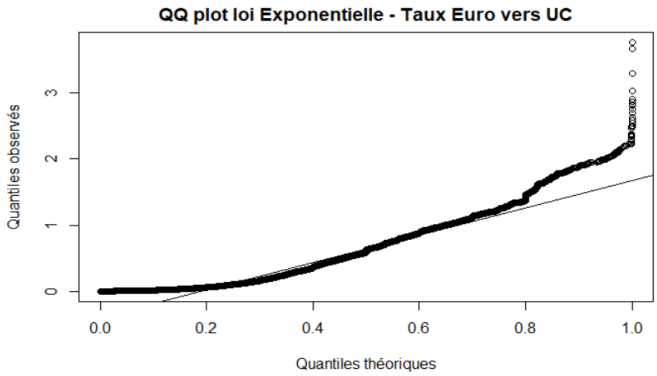

FIGURE 3.11 – QQ plot - Loi exponentielle

Le QQ-plot de la loi gamma (figure 3.9 de la page 77) est le plus satisfaisant de tous les QQ-plots. En effet, les points de ce QQ-plot sont assez bien alignés sur la droite. Il y a un décollement léger des points vers le coin en haut du graphique et également en bas du graphique. Cependant, ces décollements restent minimes devant ceux observés sur les QQ-plots des lois normale et exponentielle. Ces représentations viennent une fois de plus orienter le choix de la distribution des taux d'arbitrage vers une loi gamma. Finalement, la loi gamma sera retenue. En ce qui concerne le choix de la fonction lien, pour cette étude, la fonction lien log a été retenue car très souvent utilisée.

3.6.1.2 Choix des variables utilisées

Avant de pouvoir lancer le modèle GLM et estimer ses différents paramètres, il est impératif de faire un tri dans les variables de la base de données des mouvements d'arbitrage.

Un premier tri est effectué en regardant l'impact des variables qualitatives sur les taux d'arbitrage. Le but est de retirer du jeu de données les variables qualitatives qui n'influencent pas les taux d'arbitrage. Pour ce faire, des boîtes à moustaches ont été représentées. Les figures 3.12 et 3.13 ci-dessous sont deux exemples de représentations conduisant à des résultats différents :

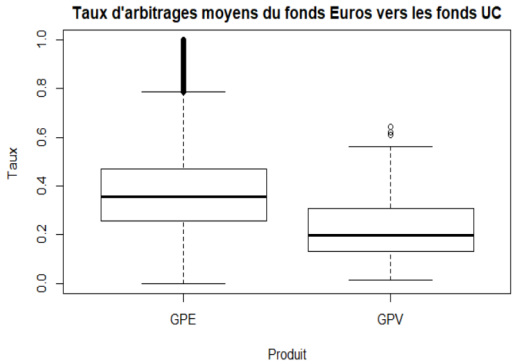

FIGURE 3.12 – Représentation des taux d'arbitrage en fonction du type de produit

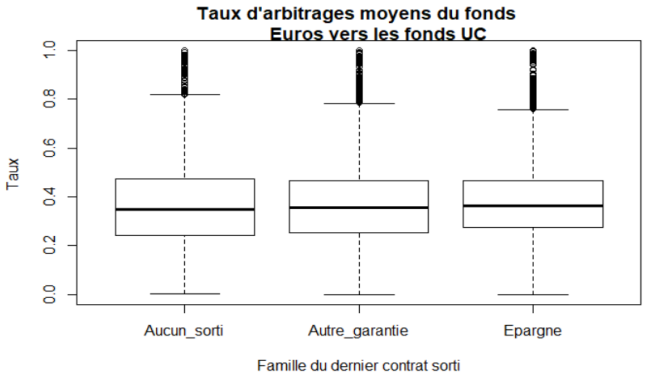

FIGURE 3.13 – Représentation des taux d'arbitrage en fonction du dernier contrat sorti

La figure 3.12 montre que le type de produit a un impact sur le taux d'arbitrage. Les clients GPV arbitrent des proportions moins importantes que les clients GPE. La variable *Produit* a à priori un impact sur le taux d'arbitrage. Elle sera donc gardée pour la suite. La figure 3.13 révèle que la variable *Famille du dernier contrat sorti* ne joue pas trop sur le taux d'arbitrage. En moyenne, les clients ont le même taux d'arbitrage, qu'ils aient sorti en dernier un contrat épargne, un contrat relatif à une autre garantie, ou même qu'ils n'aient sorti aucun contrat. De plus, les boîtes à moustache représentées sont de tailles similaires. La variable *Famille du dernier contrat sorti* ne sera pas conservée pour la suite de l'étude.

Pour ce qui est des variables quantitatives, le tri a été réalisé en regardant les différentes corrélations entre elles. Des variables trop corrélées entre elles ne peuvent pas être simultanément introduites dans les modèles.

A l'issue de tous ces traitements, les variables présentes dans la base de données sont : *Cp1AcpFi* qui représente la première composante de l'ACP réalisée avec les données financières, *Ancienneté*, *AnnéeMouvement*, *AnciennetéEnMultiEquipement*, *Produit*, *PmUC*, *PmEuro*, *PourcentageUC*.

3.6.1.3 Le GLM finalement choisi

Le nombre de variables explicatives retenues n'étant pas important, il est possible de sélectionner le meilleur modèle en ayant recours à l'approche exhaustive. Cette procédure de sélection des variables envisage de considérer tous les modèles possibles. Parmi toutes les combinaisons différentes de variables explicatives, le meilleur modèle sera celui qui a le plus faible AIC [10].

La figure 3.14 ci-dessous a aidé à faire la sélection :

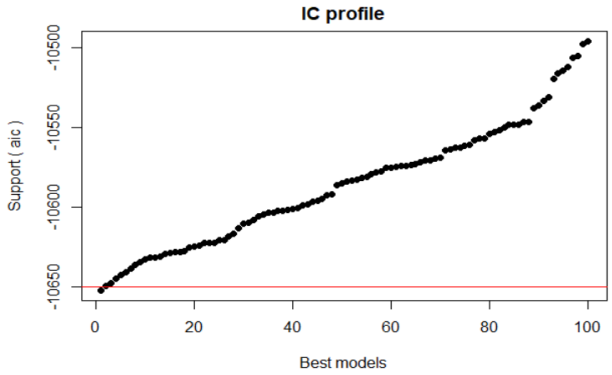

FIGURE 3.14 – Représentation des valeurs obtenues pour l'AIC des différents modèles

La ligne rouge verticale permet de démarquer les modèles dont l'AIC est inférieur de plus de 2 unités par rapport à celui du meilleur modèle. Ici, le meilleur modèle sélectionné est celui qui utilise la totalité de l'ensemble des variables. En s'appuyant également sur la figure 3.15 ci-dessous, il est décidé de finalement toutes les garder :

10. AIC = Critère d'information d'Akaike. Il existe plusieurs critères pour sélectionner $p-1$ variables explicatives parmi k variables explicatives disponibles, avec $k > p-1$. Plus d'informations sur le critère AIC et d'autres critères se trouvent en annexe

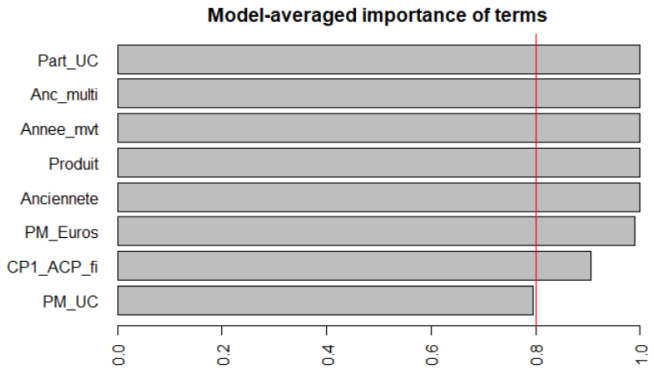

FIGURE 3.15 – Importance des différentes variables

Cette figure 3.15 met en avant l'importance de chacune des variables explicatives dans la procédure de sélection du meilleur modèle. La droite rouge verticale représente le seuil au-delà duquel une variable est supposée importante. L'importance de toutes les variables dépasse largement ce seuil, sauf dans le cas de la variable $PmUC$. Toutefois, dans la mesure où son importance est très proche du seuil, elle sera tout de même conservée pour la suite. Maintenant que tous les paramètres sont choisis, le modèle GLM va pouvoir être lancé.

3.6.1.4 Les résidus du modèle

Avant d'aller plus loin afin de commenter les résultats du modèle, il est important de regarder sa significativité globale. Pour cela, un test de significativité globale est requis. Ce dernier consiste à tester la nullité simultanée de tous les coefficients estimés par le modèle. La p-value obtenue lors de la réalisation de ce test permet d'affirmer que le modèle est globalement significatif.

Une fois la significativité globale du modèle avérée, il est temps de s'attaquer aux résidus. Leur analyse permet d'avoir une idée sur la qualité d'ajustement du modèle. Le diagnostic est essentiellement graphique. Il s'agit de regarder les figures 3.16 et 3.17 ci-dessous :

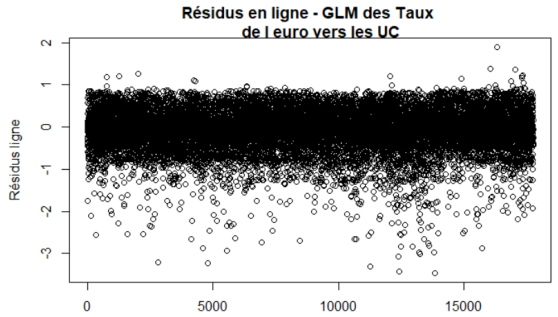

FIGURE 3.16 – Résidus ligne - Modèle GLM pour les taux d'arbitrage du fonds euros vers les fonds UC

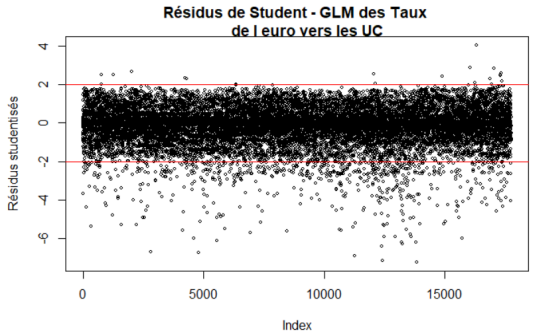

FIGURE 3.17 – Résidus de Student - Modèle GLM pour les taux d'arbitrage du fonds euros vers les fonds UC

Les résidus ligne sont pour la plupart compris entre -1 et 1. Ils pourraient être centrés. Cependant, il y a tout de même un nombre important de résidus qui se trouvent en dessous de -1. Les résidus ligne ne sont donc pas totalement satisfaisants. En ce qui concerne les résidus de Student, ils sont dit corrects lors qu'ils sont compris dans l'intervalle $[-2, 2]$. Là aussi, il y a énormément de débordements. Plusieurs points se trouvent en dehors de l'intervalle. L'analyse des résidus est très peu convaincante. Cela pouvait être prévisible, dans la mesure où les paramètres de base (notamment la distribution des taux d'arbitrage) ne correspondaient pas totalement aux observations.

Dans la suite, il s'agira de choisir au mieux les paramètres de l'arbre CART et de la forêt aléatoire afin d'optimiser leurs résultats. Pour y parvenir, les variables explicatives sélectionnées lors du précédent tri seront utilisées (à savoir *Cp1AcpFi* qui représente la première composante de l'ACP réalisée avec les données financières, *Ancienneté*, *AnnéeMouvement*, *AnciennetéEnMultiEquipement*, *Produit*, *PmUC*, *PmEuro*, *PourcentageUC*).

3.6.2 Choix des paramètres pour l'arbre CART : Arbitrages du fonds euro vers les fonds UC

Les arbres CART disposent de plusieurs paramètres. Dans un premier temps, le choix de ces paramètres a été fait à l'aide de représentations graphiques. Les graphiques 3.18 et 3.19 ci dessous, ont permis de choisir une valeur pour la profondeur maximale de l'arbre et la nombre d'individus minimum par nœud.

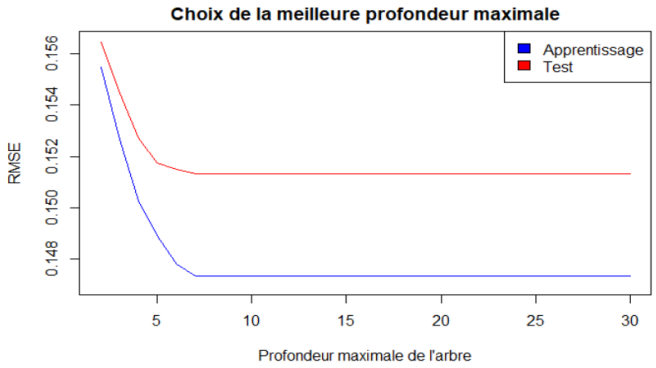

FIGURE 3.18 – Choix de la profondeur maximale

La figure 3.18 montre la décroissance du RMSE en fonction de la profondeur de l'arbre, à la fois sur le jeu de données test et le jeu de données d'apprentissage. Avant d'atteindre la profondeur 5, le RMSE décroît fortement. Ceci est le signe qu'il est avantageux d'augmenter la profondeur de l'arbre jusqu'à 5. A partir d'un profondeur maximale de 6, le RMSE devient constant. Le fait d'accroître la profondeur maximale de l'arbre n'a plus aucun intérêt. Il est donc judicieux, dans un premier temps, de retenir la valeur 5 comme profondeur maximale optimale pour l'arbre.

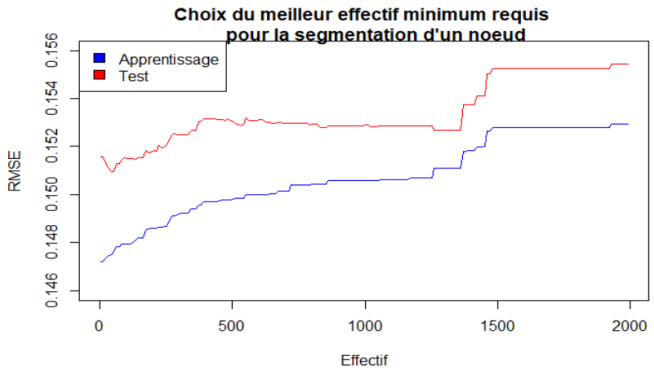

FIGURE 3.19 – Choix de l'effectif minimum requis pour diviser un nœud

Concernant la figure 3.19, le diagnostic est plus compliqué. Il y a tout d'abord une croissance progressive du RMSE en fonction de l'effectif minimum requis pour la segmentation d'un nœud (pour les deux jeux de données). Puis, cette croissance progressive laisse la place à une croissance par palliers. Il y a un premier pallier autour de 600 et au autre autour de 1500. Une possibilité serait de prendre ces valeurs pour continuer.

En fin de compte, la sélection des paramètres a été faite par tâtonnement. Les valeurs obtenues en s'appuyant sur les graphiques ont été modifiées progressivement dans l'optique d'améliorer la qualité d'ajustement du

modèle.

3.6.3 Choix des paramètres pour la forêt aléatoire : Arbitrages du fonds euro vers les fonds UC

La forêt aléatoire a été calibrée en utilisant la même méthode que celle utilisée dans le cas de l'arbre CART. Premièrement, des outils graphiques ont été utilisés pour se faire une idée sur la valeur des paramètres. Et, par la suite, ces valeurs ont été modifiées progressivement pour obtenir une meilleure version du modèle. Le graphique 3.20 ci-dessous a notamment été utilisé :

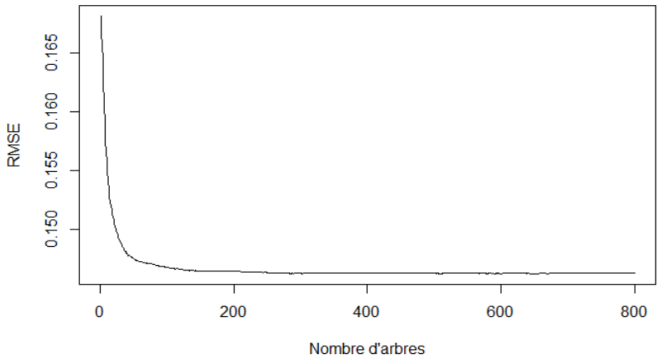

FIGURE 3.20 – Choix du nombre d'arbres de la forêt aléatoire

La qualité du modèle décroît avec le nombre d'arbres de la forêt aléatoire. Cette décroissance est dans un premier temps assez brutale, puis, elle s'observe de façon plus modérée. A partir de 200 arbres, la valeur de RMSE est plutôt constante. Cette valeur sera choisie dans un premier temps avant d'améliorer la qualité du modèle en modifiant d'autres paramètres.

3.6.4 Comparaison de tous les modèles : Arbitrages du fonds euro vers les fonds UC

Une fois tous les modèles calibrés, il est question de tous les comparer entre eux afin de sélectionner le meilleur d'entre eux. Seuls les résultats du meilleur modèle seront interprétés ensuite. Afin de pouvoir situer la qualité d'ajustement des modèles, ces derniers seront également comparés à un modèle basique. Le modèle basique ici sera celui qui prédit la moyenne de l'échantillon pour chaque observation. Pour comparer tous ces modèles, les indicateurs utilisés seront : le RMSE, le MSE, le MAE, et le MAPE [11]. Le tableau ci-dessous

11. Tous ces indicateurs sont définis en annexe

regroupe les valeurs des différents indicateurs pour tous les modèles :

Modèle	RMSE	MSE	MAE	MAPE
Modèle moyen	0,168	0,028	0,130	1,368
GLM	0,155	0,024	0,117	1,158
Arbre CART	0,151	0,022	0,114	0,822
Forêt aléatoire	0,148	0,022	0,112	0,842

TABLE 3.2 – Différents indicateurs calculés pour les modèles

Ce tableau révèle que les modèles réalisés ont une meilleure qualité de prédiction que le modèle moyen qui est trivial. Même si les écarts sont relativement faibles, la forêt aléatoire semble se démarquer des autres modèles. Elle obtient des valeurs plus faibles sur les indicateurs, signe qu'elle a une meilleure qualité d'ajustement. C'est donc la forêt aléatoire qui sera retenue.

En terme de taux prédits, la figure 3.21ci- dessous donne plus d'indications :

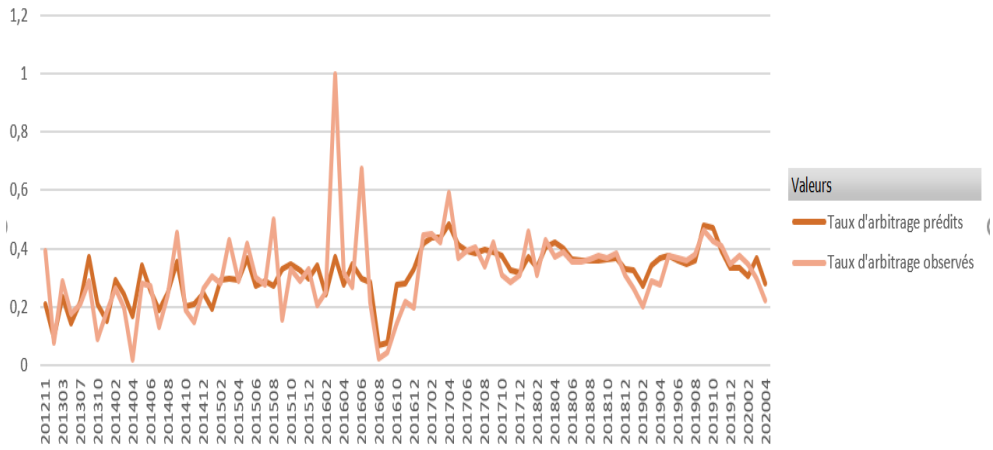

FIGURE 3.21 – Comparaison des taux d'arbitrage moyen - échantillon test

Globalement, en moyenne le modèle se rapproche assez bien des données observées. Le modèle capte bien les petites variations des taux observés et s'ajuste bien lorsqu'il y a de petits pics. En revanche, pour l'année 2016, le modèle réalise de moins bonnes performances. Cette année 2016 était déjà ressortie lors de l'analyse exploratoire des données. En effet, elle ne suivait pas la tendance observée lors des autres années. Les données observées présentent un pic important à la fin de l'année 2015, puis, une baisse rapide pendant l'année 2016. Alors qu'en parallèle, le modèle fait abstraction du pic qu'il y a à la fin de l'année 2015. Il se contente de marquer une légère baisse des taux moyens d'arbitrage pour l'année 2016.

4. Identification des profils d'assurés avantageux de ce portefeuille

Les arbitrages peuvent modifier la composition du portefeuille de l'assureur. Des arbitrages massifs vers les fonds UC sont préférables du point de vue de l'assureur, car ils permettent d'augmenter la quantité d'encours sur les fonds UC et de réduire celle sur les fonds en euros. Après une présentation des atouts des fonds en unités de compte pour l'assureur, il s'agira de mettre en lumière les caractéristiques des assurés du portefeuille étudié qui arbitrent en majorité vers les fonds UC. Le but est d'essayer de dégager des profils types. Pour ce faire, l'algorithme des *k-means* est utilisé.

4.1 Les atouts des fonds en unités de compte : Évolution du résultat pour l'assureur sur un contrat d'épargne

Comme précisé tout au long de cette étude, les contrats d'épargne de ce portefeuille offrent plusieurs types de fonds d'investissements aux clients dont les fonds en euros et les fonds UC. Le but de cette section est de voir l'impact des différents types de supports sur le résultat de l'assureur, et de comprendre au moyen d'exemples basiques et simples en quoi les fonds UC constituent un réel avantage pour les assureurs.

Les contrats d'épargne sont rythmés par plusieurs mouvements tout au long de leur durée de vie : les rachats (totaux ou partiels), les versements, les décès, les revalorisations, et bien évidemment les arbitrages. Pour des raisons de simplicité, il ne sera pas fait mention de tous ces mouvements lors de ces exemples. Ces exemples, visent à mettre simplement en avant l'impact des fluctuations des marchés financiers sur le résultat de l'assureur, en fonction du type de fonds (euro ou UC) présents sur le contrat d'épargne. Il s'agira de regarder le résultat de l'assureur dans plusieurs cas de figure : quand les marchés financiers sont favorables (et donc que les taux de rendements sont élevés), et ensuite lorsque les marchés financiers sont un peu moins rentables (et que les taux servis sont plutôt bas).

4.1.1 Exemple 1 : Le cas d'un contrat épargne 100 % euros

Cet exemple se base sur le cas d'un contrat épargne 100 % euros. L'épargne du client est supposée investie en totalité sur un seul fond euros. Les autres hypothèses du contrat sont les suivantes :

- Clause de PB contractuelle : 90 % ;
- TMG : 1 % ;
- Frais de gestion : 0,8 %.

Le tableau 4.1.1 de la page 86 ci-après représente la marge possible pour l'assureur dans plusieurs cas de figure pour ce contrat :

Situation	**Taux de rendement brut des actifs**	**Taux net servi**	**Marge assureur**
Marchés financiers favorables	3%	$1,9\%$	$1,1\%$
Marchés financiers en baisse	$-1,5\%$	1%	$-2,5\%$

Table 4.1 – L'impact des fluctuations des marchés financiers sur les contrats d'épargne : Cas des fonds en euros

Le résultat de l'assureur est directement lié aux fluctuations des marchés financiers. Si les rendements sur le marché sont favorables, la marge de l'assureur est bonne. En cas de baisse des rendements des actifs, le résultat de l'assureur est directement impacté. Cet effet est amplifié par les différentes options sur le contrat. La présence du TMG creuse d'autant plus la perte de l'assureur dans le cas où les marchés financiers sont en baisse. Elle serait de $-1,5\%$ sans TMG. L'assureur s'engage à fournir au client un résultat positif sur le fonds euros (résultat au moins égal au TMG s'il y en a, ou au moins nul sinon). De plus, le prélèvement des frais de gestion devient presque impossible dans ce genre de situation. L'assureur est alors obligé de puiser dans ses réserves pour rééquilibrer la balance.

4.1.2 Exemple 2 : Le cas d'un contrat épargne 100 % UC

Cet exemple se base sur le cas d'un contrat épargne 100 % UC. L'épargne du client est supposée investie en totalité sur un seul fonds UC. Les autres hypothèses du contrat sont les suivantes :

- Capital investi à l'entrée : 100 000 euros ;
- Valeur de part : 100 euros ;
- Taux de frais de gestion : 1 %.

Compte tenu de ces hypothèses, le client dispose sur son contrat de 1 000 parts d'UC. Il s'agit maintenant de regarder l'impact de la modification de la valeur de part (liée directement aux fluctuations des marchés financiers) sur la marge de l'assureur grâce au tableau 4.1.2 ci-dessous :

Situation	**Valeur de part**	**Valeur encours**	**Marge assureur**
Hausse de la valeur de part	120	120 000	1 200
Baisse de la valeur de part	80	80 000	800

Table 4.2 – L'impact des fluctuations des marchés financiers sur les contrats d'épargne : Cas des fonds UC

Sur les fonds UC, l'assureur utilise les frais de gestion pour se rémunérer. Ils représentent un pourcentage (ici égal à 1%) de l'encours de l'UC. Peu importe le sens de variation de la valeur de part, la marge dégagée par l'assureur est toujours positive. Seul le client subit les fluctuations des valeurs de parts car le montant de son épargne y est directement lié. Certes le résultat de l'assureur baisse lorsque les valeurs de part baissent, mais son résultat reste tout de même positif. La prise de frais reste toujours possible, peu importe la performance de l'UC. L'assureur peut ainsi continuer à générer des bénéfices même en situation défavorable. L'engagement de l'assureur concernant ce genre de support est donc moindre. Ce type de fonds ne nécessite pas de puiser dans les réserves de l'assureur, tant qu'il reste en capacité d'absorber ses coûts fixes.

Pour la suite, les profils des assurés qui arbitrent fréquemment dans le sens des fonds en unités de compte seront précisés grâce notamment à l'algorithme des *k-means*.

4.2 Principe de l'algorithme des *k-means*

L'algorithme des *k-means* est un algorithme de clustering. Il s'applique sur les variables quantitatives uniquement. Il s'agit d'une routine permettant d'aborder un autre type de problème de clustering. Ici, le nombre de classes à former est fixé à k. Les k groupes d'individus (créés en minimisant la distance entre les points à l'intérieur de chaque partition) sont distincts et aussi homogènes que possible.

Etant donnée une partition de départ en k groupes, il faut déterminer pour chaque groupe son centre de gravité ; puis reformer les groupes en associant ensemble les points qui sont les plus proches d'un centre de gravité. La procédure est itérée jusqu'à satisfaction d'un critère d'arrêt (généralement la stabilisation des groupes). Il faut bien comprendre qu'à l'issue des itérations, le but est d'obtenir une partition de bonne qualité. Cependant, il n'y a aucune garantie d'optimalité globale (par rapport à l'inertie intraclasse). Le pseudo code ci-dessous représente l'algorithme :

Entrée	Le nombre de classes k à former, le jeu de données à partitionner
Initialisation	Choisir k points $(C_1^{(1)}, \ldots, C_k^{(1)})$ qui seront les centres de gravité des clusters (dits centroïdes)
Itération	Affecter chaque point au groupe dont il est le plus proche par rapport à son centre
	C'est-à-dire, constituer une partition $S_i^{(t)}$ telle que :
	$S_i^{(t)} = \{x_j \text{ tel que } \|\|x_j - C_i^{(t)}\|\| \leq \|\|x_j - C_{i^*}^{(t)}\|\|, \forall i^* = 1, \ldots, k\}$
	Recalculer le centre de chaque cluster et modifier le centroïde $m_i^{t+1} = \frac{1}{\|S_i^{(t)}\|} \sum_{x_j \in S_i^{(t)}} x_j$
Condition d'arrêt	Convergence (Centroïdes qui ne changent plus ou nombre d'itérations maximum atteint)
	Ou stabilisation de l'inertie totale de la population
Sortie	Les k groupes partitionnés

TABLE 4.3 – L'algorithme des *k-means* expliqué

Pour mieux comprendre l'algorithme, la figure 4.1 ci-dessous peut être utilisée. Il s'agit des différentes étapes de l'algorithme illustrées sur un jeu de données. Au cours de cet exemple [13], le but est de séparer les données en deux groupes homogènes :

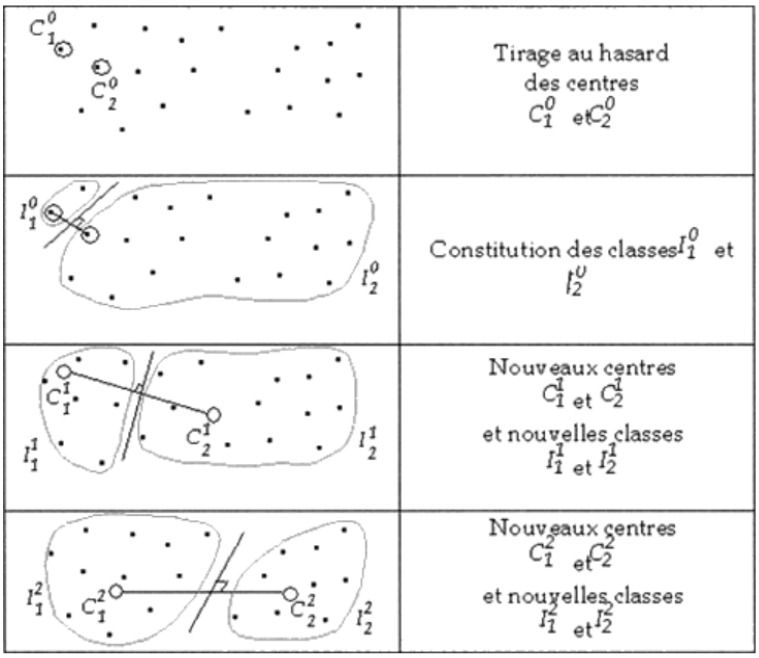

FIGURE 4.1 – L'algorithme des k-means illustré

Il s'agit d'un algorithme particulièrement simple et facile à comprendre. Il faut toutefois faire attention. Le résultat de l'algorithme est fortement lié au choix initial des centroïdes et à la distance utilisée. En fonction des centres de gravité initiaux, les classes obtenues sont différentes. De ce fait, la configuration des clusters trouvés par l'algorithme peut ne pas être la plus optimale (cas des optimum locaux). Par ailleurs, le nombre k de classes sélectionné au départ joue un rôle très important. En effet, le choix de k conditionne le résultat final. Une stratégie pour identifier le nombre de classes consiste à faire varier k et surveiller l'évolution de l'inertie intra-classes. L'idée est de visualiser le coude où l'adjonction d'une classe ne correspond à rien dans la structuration des données. Il faut sélectionner k en fonction de la dernière classe supplémentaire à induire un gain informationnel significatif.

4.3 Résultats

Pour appliquer l'algorithme des *k-means* au portefeuille étudié, la fonction *kmeans* du logiciel R a été utilisée. Cette partie est consacrée uniquement aux clients de la base de données qui ont effectué au moins 3 arbitrages du fonds euro vers les fonds UC au total pendant toute la durée d'étude. Il y a très peu de clients qui osent effectuer plus de 3 arbitrages. Ces clients sont très intéressants pour l'assureur, qui a tout intérêt à les garder dans son portefeuille. Pour classer ces assurés, il s'agit de s'intéresser aux variables qui ont été retenues lors de la mise en place de nos modèles dans la partie précédente. Par la suite, toutes les autres variables (notamment celles qui n'ont pas été retenues pour la modélisation des taux d'arbitrage) seront aussi mises en valeurs.

4.3.1 La classification

Avant de regarder leurs profils, il peut être judicieux de regarder le nombre d'arbitrages qu'ils ont actés en fonction des différentes années. La figure 4.2 ci-après montre le nombre d'arbitrages effectués par an pour les assurés qui arbitrent énormément :

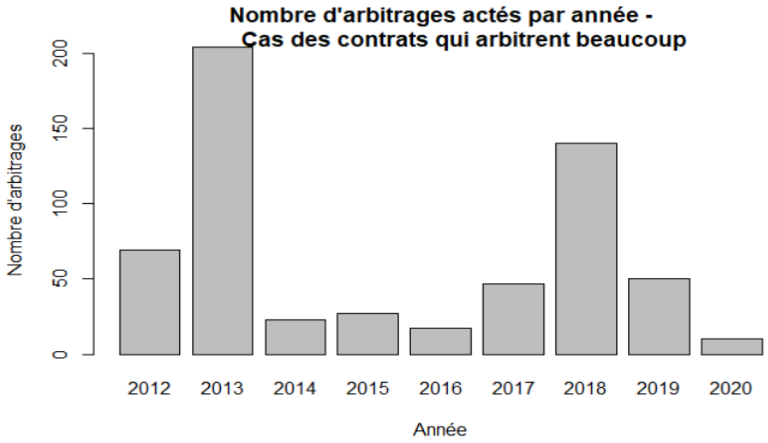

FIGURE 4.2 – Nombres d'arbitrages par année - Cas des contrats qui arbitrent beaucoup

Les "grands arbitreurs" sont très réactifs pendant les années 2013 et 2018. Pendant ces années, le nombre d'arbitrages effectués dépasse 120. Ils ont été très réceptifs lors de la première campagne d'arbitrage. Mais, contrairement aux autres assurés jugés "opportunistes" qui se sont contentés d'arbitrer uniquement pendant les campagnes de *business transformation*, ces assurés font vivre leur contrat d'épargne aussi en temps normal.

La première étape avant l'étude des caractéristiques de ces assurés est la réalisation d'une ACP avec les différentes variables explicatives de la base de données. Cette ACP sera utile pour la représentation conjointe des individus et des variables explicatives. Cela permettra une interprétation plus pertinente des résultats. Pour l'ACP, 3 composantes principales ont été conservées. Elles permettaient d'expliquer un pourcentage correct de l'inertie totale, à savoir 68%.

Ensuite, il est temps de se décider sur le nombre de classes d'assurés à faire. Les figures 4.3 et 4.4 ci-dessous permettent d'aguiller ce choix :

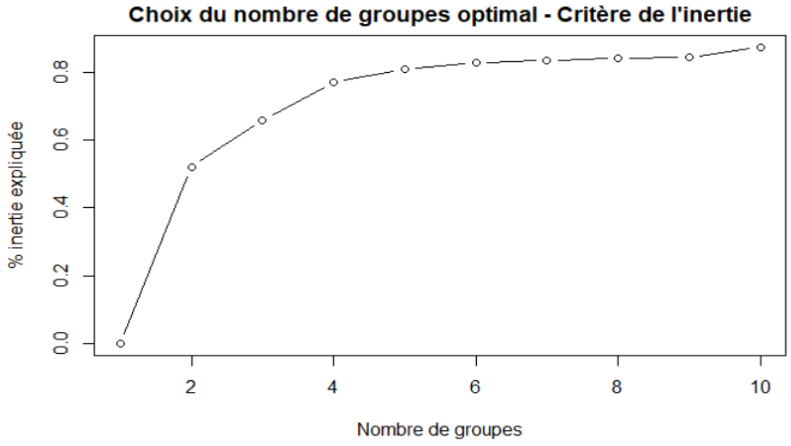

FIGURE 4.3 – Pourcentage d'inertie expliquée en fonction du nombre de classes

La figure 4.3 met en lumière le fait qu'à partir de 4 classes, l'adjonction d'un groupe supplémentaire n'augmente pas significativement la part d'inertie expliquée par la partition. En se basant sur cette figure, il faudrait ainsi partitionner en 3 classes.

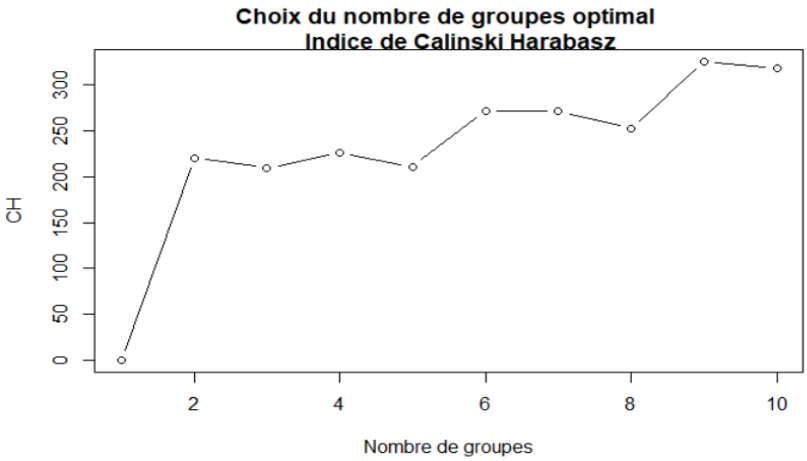

FIGURE 4.4 – Nombres d'arbitrages par année - Cas des contrats qui arbitrent beaucoup

En revanche, la figure 4.4 oriente davantage le choix vers la segmentation en 2 groupes. L'indice de Calinski-Harabasz [1] est maximal lors de la séparation en 10 groupes. Cependant, dans un souci de parcimonie, ce nombre de groupes ne sera pas retenu. À partir de deux classes, le critère de Calinski-Harabasz augmente graduellement par paliers. Finalement, au regard du pourcentage d'inertie expliqué et de la valeur du critère de Calinski-Harabasz, 2 groupes seront retenus.

1. L'indice de Calinski-Harabasz est une mesure de qualité d'une partition d'un ensemble de données en classification automatique. C'est le rapport entre la variance intergroupes et la variance intra-groupe. Plus il est élevé, plus la partition est meilleure.

Les caractéristiques des deux groupes d'assurés sont données par les graphiques 4.5 et 4.6 ci-dessous :

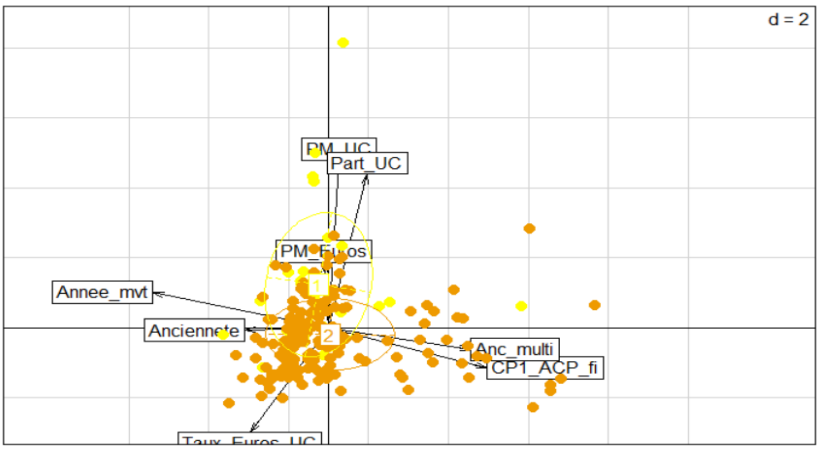

FIGURE 4.5 – Représentation des groupes d'assurés dans le premier plan factoriel

Le premier plan factoriel est caractérisé par les variables *PmUC*, *PartUC*, *AnciennetéEnMultiEquipement*, et *Cp1AcpFi*. Le premier groupe en orange n'est pas suffisamment bien représenté pour permettre d'en donner tous les attributs. Cependant, il y a quelques individus de ce groupe qui se démarquent. Il serait possible qu'ils réalisent leurs arbitrages lorsque les valeurs sur le marché financier sont importantes. De plus, il semblerait qu'ils aient basculé en multi-équipement depuis longtemps. Les assurés du groupe jaune, quant à eux, se distingueraient potentiellement par leur épargne UC et leur pourcentage d'UC élevés. Il s'agit de clients "riches" à priori.

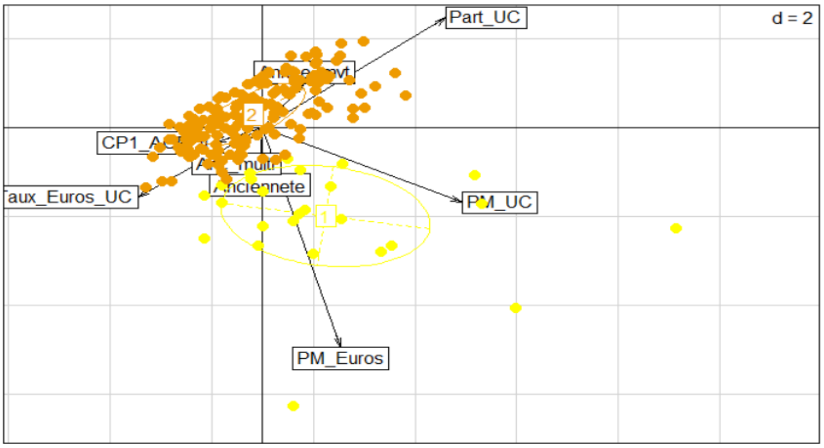

FIGURE 4.6 – Représentation des groupes d'assurés dans le deuxième plan factoriel

Le deuxième plan factoriel est le plan de la richesse. Il se caractérise par les variables *PmEuro*, *PmUC*. Cette représentation révèle que les assurés du groupe jaune se distinguent par une épargne assez importante.

En résumé, les deux groupes d'assurés formés sont assez différents. Il y a un premier groupe, qui a une forte appétence pour le risque. Ce groupe d'assurés dispose de contrats d'épargne avec un fort pourcentage d'UC et un gros montant investi sur les UC. L'autre groupe est composé d'assurés réactifs aux variations sur les marchés financiers. Cela laisse penser qu'ils consultent souvent les marchés financiers et attendent le moment opportun pour effectuer leurs arbitrages. Ces assurés seraient peut-être plus attentifs à leur contrat d'épargne car ils détiennent un nombre important de contrats depuis longtemps.

4.3.2 Leurs caractéristiques

D'autres caractéristiques ressortent après l'analyse des deux groupes :

— La richesse :

  Les assurés du groupe jaune sont moins nombreux que ceux du groupe orange. Ils sont seulement au nombre de 27, contre 178 pour le groupe jaune. Comme l'algorithme de *k-means* l'a révélé, ces assurés sont aisés. Le moins riche d'entre eux compte 165 000 d'encours au total. En moyenne, ces assurés ont 328 000 euros d'encours. Ceci n'est pas surprenant vu qu'il s'agit de clients GPE. Pour ce qui est des assurés du groupe orange, l'encours moyen observé dépasse à peine 3 000 euros. Pour ce groupe, la moyenne de l'encours est relativement plus basse. Elle est d'environ 55 000 euros. Ce groupe concerne à la fois des assurés GPE et GPV. L'analyse du montant de l'encours confirme donc ce que la classification laissait déjà apparaître.

— Leur catégorie socio-professionnelle :

  Il ressort que les assurés des deux groupes comptent en majorité des retraités. Le groupe orange recense également une bonne partie d'ouvriers et d'employés divers.

— Les taux d'arbitrage pour les arbitrages du fonds euro vers le fonds UC :

  En moyenne, le taux d'arbitrage ne permet pas forcément de distinguer les assurés des deux groupes. Pour les deux groupes, le taux moyen d'arbitrage est autour de 30 %.

— Le nombre de contrat détenus :

  Le nombre de contrats détenus en moyenne par les assurés du groupe orange (5 contrats) est plus élevé que celui des assurés du groupe jaune (3 contrats). La quasi-totalité des assurés du groupe jaune se limitent à 3 contrats.

5. Conclusion et Limites

Ce mémoire est axé sur la modélisation des mouvements d'arbitrage d'un portefeuille ayant bénéficié de deux campagnes d'arbitrage. Ces campagnes de *business transformation* portent bien leur nom. Elles représentent un levier qui permet à l'assureur de transformer la composition de son portefeuille à son avantage.
L'analyse des volumes arbitrés révèle que hormis pendant les campagnes, les clients de ce portefeuille sont très peu réactifs en terme d'arbitrages. Cependant, Generali a tout intérêt à garder en mémoire les profils des groupes assurés qui ont bien joué le jeu et qui arbitrent plus souvent que la moyenne. Il s'agit soit de clients riches qui ont un encours important sur leur contrat, soit, de clients qui disposent de plusieurs contrats.

Les principaux éléments qui influencent les arbitrages sont : le type de produit (GPE ou GPV), l'ancienneté du contrat, le taux de frais appliqué au moment de l'arbitrage, le sexe de l'assuré et le nombre de contrats détenus. Les clients GPE, plus riches en moyenne, arbitrent des montants moyens plus importants que les clients GPV. Avec l'ancienneté, les clients sont plus à l'aise et ont tendance à arbitrer des montants moyens plus importants jusqu'à 9 ans d'ancienneté. Après 9 ans d'ancienneté, la tendance s'inverse et les montants moyens arbitrés baissent. Pour ce qui est des taux de frais d'arbitrage, il apparaît que les clients sont plus à même d'effectuer des arbitrages lorsque les taux de frais d'arbitrage sont bas. En revanche, il semblerait que le sexe de l'assuré n'ait pas une grande incidence sur les montants arbitrés ou la fréquence d'arbitrage. Les hommes arbitrent autant que les femmes et dans les mêmes proportions en moyenne.

Concernant les modèles utilisés pour les mouvements d'arbitrages, les forêts aléatoires se sont avérées être les plus robustes, dans le cas des arbitrages du fonds euro vers le fonds UC, mais, également dans le cas des arbitrages des fonds en UC vers le fonds euro. Les forêts aléatoires présentent néanmoins un problème de lisibilité. Elles sont rapidement confrontées à des problèmes d'interprétation. Il est difficile d'expliquer les résultats : c'est l'effet boîte noire.

L'étude réalisée peut trouver son utilité dans le cadre du calcul des provisions de fin de période du portefeuille considéré. Elle doit toutefois être complétée. Il aurait été possible, dans un premier temps, de regarder la globalité du portefeuille (et pas seulement les clients qui ont acté des arbitrages comme dans le cas de cette étude), afin de modéliser leur décision d'arbitrage. De plus, à la modélisation des mouvements d'arbitrages, il faudrait rajouter celle des autres mouvements types d'un contrat d'épargne : les rachats, les versements par exemple.

Generali a réfléchi à d'autres astuces pour réduire le risque afférent à ses fonds en euros. Ces actions peuvent se faire en amont, au moment de la collecte, ou alors directement sur le stock déjà présent dans le portefeuille. Par exemple, Generali propose déjà des contrats d'épargne avec un capital sur les fonds en euros garanti "brut de frais de gestion". Jusqu'à maintenant, le capital versé sur les fonds en euros était garanti net de frais de gestion. Dans ce cas de figure, l'assuré est sûr de récupérer au minimum la totalité du capital investi. Avec les fonds en euros garantis "brut de frais de gestion", le capital garanti baisse chaque année du fait du prélèvement des frais de gestion sur le contrat.

Annexes

1. Quelques indicateurs de prédiction

Il existe plusieurs indicateurs qui permettent de vérifier la qualité de prédiction d'un modèle quantitatif. Ici, tous ces indicateurs sont calculés à partir du jeu de données test.

Dans toute cette partie :

- n représente la taille du jeu de données de test ;
- y_i est la i^{me} observation du jeu de données de test ;
- Et $\hat{y}_i$ est la valeur prédite pour la i^{me} observation par le modèle.

1 RMSE

La racine carré de l'erreur quadratique moyenne (**RMSE** pour Root Mean Square Error en anglais) se définit comme suit :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{1.1}$$

2 MSE

L'erreur quadratique moyenne (**MSE** pour Mean Square Error en anglais) se définit comme suit :

$$MSE = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n} \tag{1.2}$$

3 MAE

L'erreur moyenne absolue (**MAE** pour Mean Absolute Error en anglais) se définit comme suit :

$$MAE = \sqrt{\frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}} \tag{1.3}$$

4 MAPE

La moyenne de la valeur absolue des pourcentages d'erreur (**MAPE** pour Mean Absolute Percentage Error en anglais) se définit comme suit :

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{\hat{y}_i - y_i}{y_i}| \tag{1.4}$$

5 La valeur en risque (VaR - Value at risk)

Selon Engle et Manganelli [10], la VaR peut être définie comme la perte maximale potentielle qui ne devrait être atteinte qu'avec une probabilité donnée sur un horizon temporel donné. En considérant un taux de couverture α (autrement dit un niveau de confiance $1-\alpha$), la VaR correspond au quantile de niveau α de la distribution de perte et profit valable sur la période de détention de l'actif. Plus le niveau de couverture du risque de ruine est élevé, plus le besoin en capital sera important.

$$VaR(\alpha) = F^{-1}(\alpha)$$

Avec $F(.)$ la fonction de répartition associée à la distribution de perte et profit.

La VaR est très utilisée dans le domaine de la finance pour mesurer par exemple le risque de marché afférent à un portefeuille d'instruments financiers. Dans le cadre du SCR, la VaR de niveau α définit le capital seuil, tel que, un besoin supérieur en capital ait une probabilité $1-\alpha$ de se réaliser avec $\alpha = 0,5\%$. La VaR peut aussi être utilisée en assurance lors du calcul de provision.

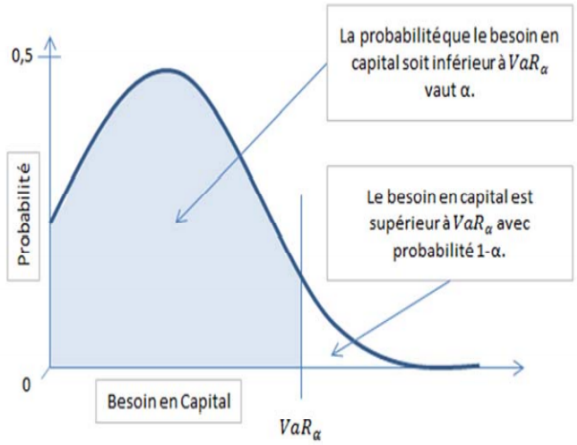

FIGURE 1.1 – La VaR illustrée au moyen d'une courbe en besoin de capital

2. Lexique

- UC : Unités de compte
- UCT : Unité de compte tiers
- UCM : Unités de compte maison
- PU : Prime unique
- PP : Prime périodique
- RC : Responsabilité civile
- PNO : Propriétaire non-occupant
- IARD : Incendies accidents et risques divers
- MRH : Multi risques habitation
- MRC : Multi risques construction
- GAV : Garantie accidents de la vie
- AGIRC : Association générale des institutions de retraite complémentaire des cadres
- ARRCO : Association pour le régime de retraite complémentaire des salariés
- GPV : Generali Protection Vie
- TMA : Taux minimum annoncé

3. Pour compléter le modèle GLM

1 Les critères de sélection des variables

Les informations ci-dessous sont tirées de différents ouvrages [1] que le lecteur peut retrouver dans le bibliographie.

Il existe plusieurs critères pour sélectionner $p-1$ variables explicatives parmi k variables explicatives disponibles, avec $k > p-1$. Ici, le lecteur pourra trouver la définition de certains de ces critères.

1.1 Le critère de R^2

Le R^2, dit coefficient de détermination est un indicateur qui permet de jauger la qualité d'un modèle linéaire. Il est donné par la formule ci-dessous :

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}$$

Avec n le nombre d'observations, y_i la valeur de la i^{eme} observation, $\hat{y}_i$ la valeur prédite par le modèle pour la i^{eme} observation, et $\bar{y}$ la moyenne des observations.

La valeur du R^2 est comprise entre 0 et 1. Plus il est proche de 1, plus le modèle est performant et est capable d'expliquer une grande partie des observations. Dans le cas contraire, un R^2 proche de 0 traduira une mauvaise qualité d'ajustement du modèle.

Le R^2 a un inconvénient majeur. Il a tendance à augmenter de façon monotone avec l'introduction de nouvelles variables même si celles-ci sont peu corrélées avec la variable expliquée. Il est parfois nécessaire de se tourner vers d'autres indicateurs.

1. Consulter [4] et [6]

1.2 Le critère du R^2_{ajuste}

Le R^2_{ajuste} est principalement utilisé pour comparer des modèles ayant un nombre de variables explicatives différent. Il permet de faire un choix plus aiguisé que le R^2 classique. Entre deux modèles, le meilleur modèle est celui avec la plus grande valeur de R^2_{ajuste}. Le R^2_{ajuste} est donné par la formule ci-dessous :

$$R^2_{ajuste} = 1 - \frac{n-1}{n-p-1} \times (1 - R^2) = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2/(n-p-1)}{\sum_i^n (y_i - \bar{y})^2/(n-1)}$$

Attention, le R^2_{ajuste} peut prendre des valeurs négatives.

1.3 La statistique du C_p Mallows

Le C_p Mallows est un indicateur introduit par Colin Lingwood Mallows. Entre deux modèles, le meilleur modèle sera celui qui aura la plus petite valeur de C_p. Dans le cas où le nombre total de variables explicatives disponible est k, un modèle sélectionnant uniquement p variables aura un C_p donné par :

$$C_p = \frac{\sum_i^n (y_i - y_{pi})^2}{S^2} - n + 2(p+1)$$

Avec :

- y_{pi} : La valeur prédite pour la i^{eme} observation par le modèle ;
- n : Le nombre total d'observations ;
- S^2 : Le carré moyen résiduel après régression sur l'ensemble complet de k régresseurs.

Une des limites du C_p est qu'il ne peut être utilisé que pour des modèles qui disposent d'un échantillon de grande taille.

1.4 Le critère AIC

Le critère d'information d'Akaike dit AIC a été mis en place par Hirotugu Akaike en 1973. Il est donné par la formule ci-dessous :

$$AIC = 2k - 2\ln(L)$$

Avec k le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.

Entre deux modèles candidats, le meilleur modèle sera celui qui aura la plus petite valeur d'AIC. L'AIC représente un compromis entre le biais et la parcimonie. Il pénalise les modèles avec un nombre important

de paramètres. Dans le cas où le nombre de paramètres k est grand devant n la taille de l'échantillon, il est préférable de se tourner vers l'AIC_c.

1.5 Le critère AIC_c

L'AIC_c est une version améliorée de l'AIC classique qui s'utilise dan sle cas des échantillons de petite taille. Il est donné par la formule ci-dessous :

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

Où n représente la taille de l'échantillon et k le nombre de paramètres à estimer.

1.6 Le critère du BIC

Le BIC est au autre critère inspiré de l'AIC. Il est donné par la formule suivante :

$$BIC = -2\ln(L) + \ln(n)k$$

Où n représente la taille de l'échantillon et k le nombre de paramètres à estimer.

Le BIC se trouve être plus parcimonieux que l'AIC. En effet, il pénalise davantage le nombre de variables présentent dans le modèle.

2 L'algorithme de Newton-Raphson

Pour trouver le zero d'une fonction, l'algorithme de Newton-Raphson utilise la dérivée. L'idée consiste à partir d'une approximation quelconque x_0 du zéro d'une fonction f et d'approcher le graphe de f au voisinage de x_0 par celui de sa tangente T_0 au même point $(x_0, f(x_0))$. La tangente T_0 a pour équation :

$$y = f'(x_0)(x - x_0) + f(x_0)$$

Cette tangente coupe l'axe des abscisses en un point de coordonnées $(x_1, 0)$. Dans le cas où $f'(x_0)$ est non nul, il est possible d'écrire :

$$x_1 = x_0 + \frac{f(x_0)}{f'(x_0)}$$

En itérant le processus (tant que cela est possible, c'est-à-dire tant que la dérivée est non nulle au point considéré), il est possible de construire la suite suivante : $x_{n+1} = x_n + \dfrac{f(x_n)}{f'(x_n)}$. La suite (x_n) converge vers le zéro de la fonction. La figure 3.1 illustre cet algorithme :

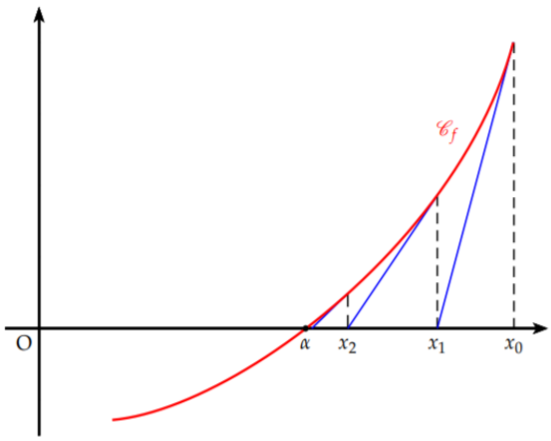

FIGURE 3.1 – Algorithme de Newton-Raphson : La suite (x_n) converge vers le point α

4. Le calibrage des différents modèles : Arbitrages des fonds UC vers le fonds euro

Dans cette section, le lecteur pourra retrouver toutes les figures ayant aidé au choix des paramètres des modèles pour les arbitrages des fonds UC vers le fonds euro. Les taux moyens par année des arbitrages des fonds UC vers le fonds euro se trouvent dans la figure ci-dessous :

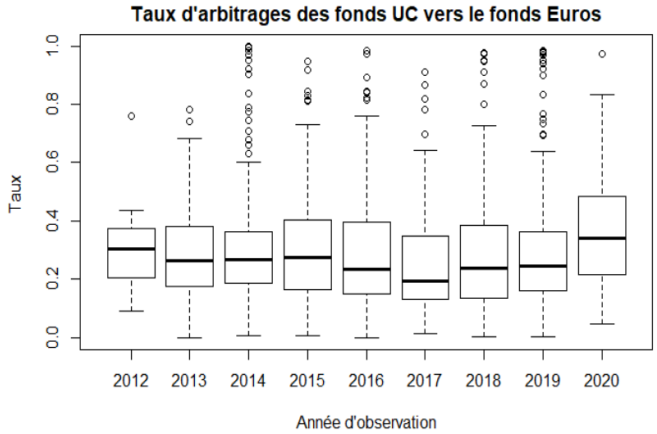

FIGURE 4.1 – Les taux moyens d'arbitrages par année : Arbitrages des fonds UC vers le fonds euro

1 Le modèle GLM : Arbitrages des fonds UC vers le fonds euro

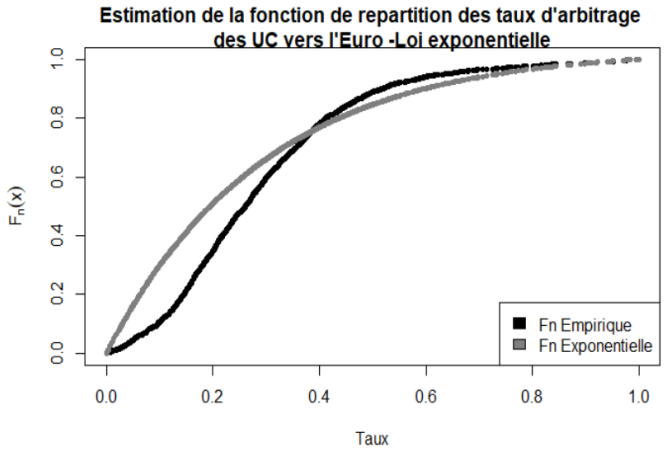

FIGURE 4.2 – Comparaison des fonctions de répartition : Cas de la loi exponentielle

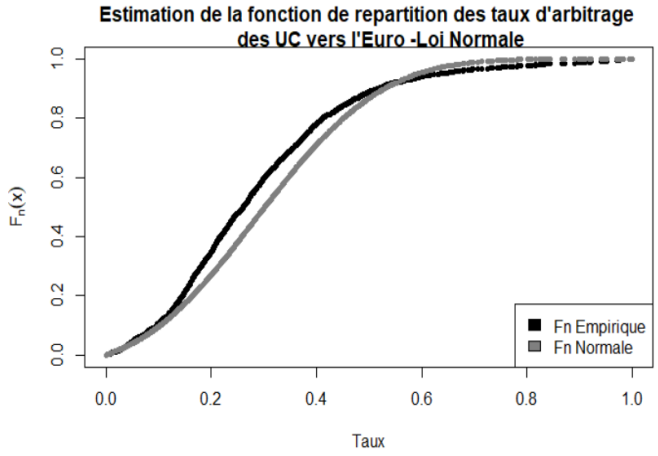

FIGURE 4.3 – Comparaison des fonctions de répartition : Cas de la loi normale

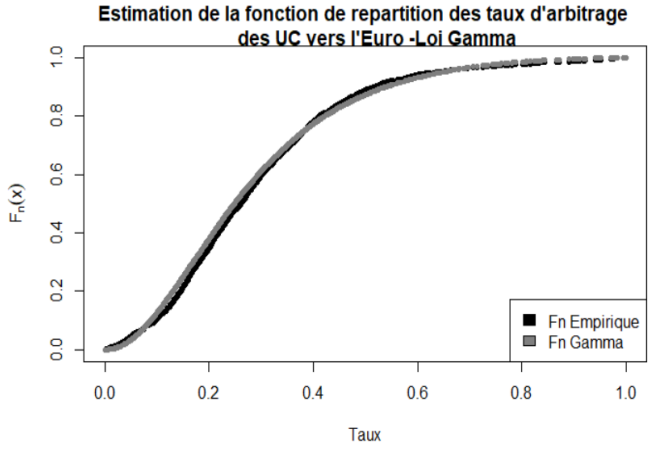

FIGURE 4.4 – Comparaison des fonctions de répartition : Cas de la loi gamma

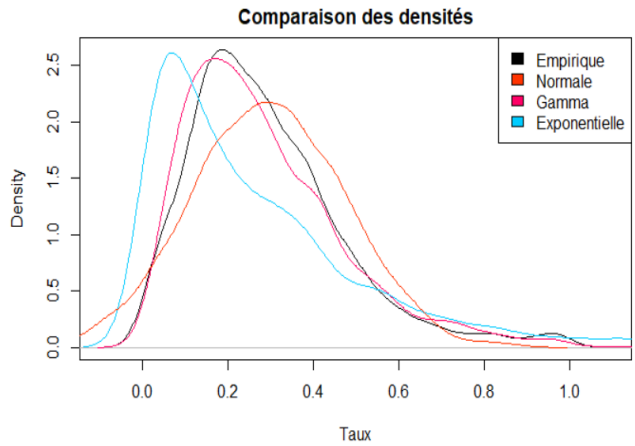

FIGURE 4.5 – Comparaison des densités

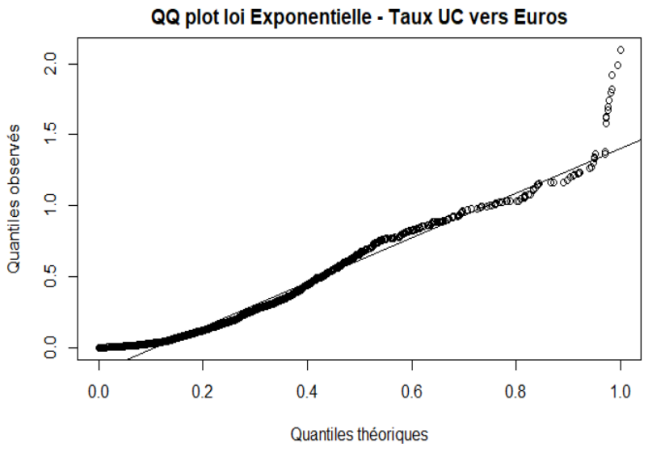

FIGURE 4.6 – QQ-plot : Cas de la loi exponentielle

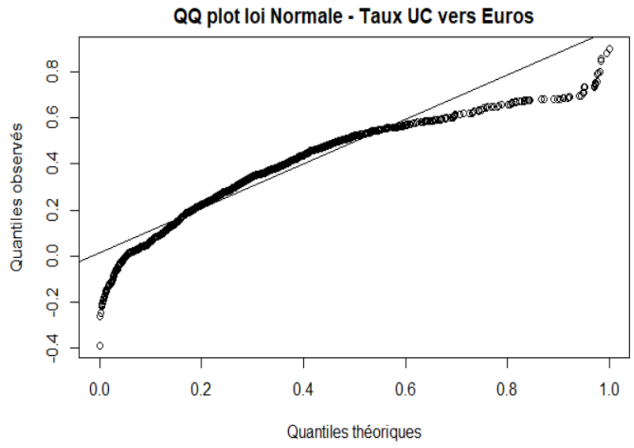

FIGURE 4.7 – QQ-plot : Cas de la loi normale

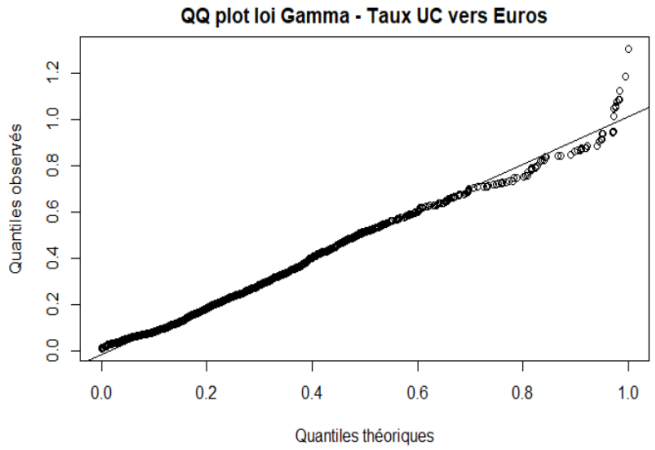

FIGURE 4.8 – QQ-plot : Cas de la loi gamma

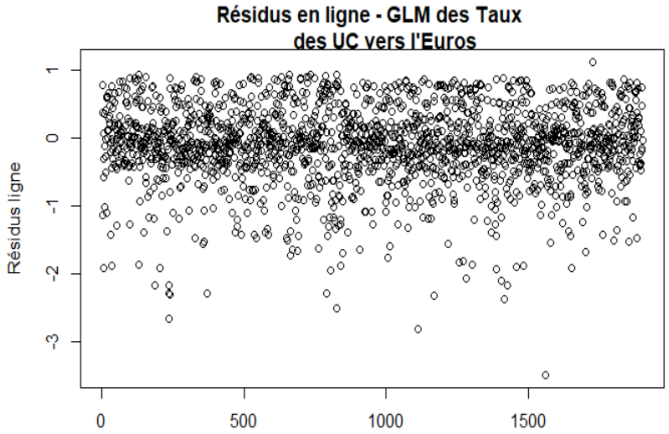

FIGURE 4.9 – Résidus ligne : Arbitrages des fonds UC vers le fonds euro

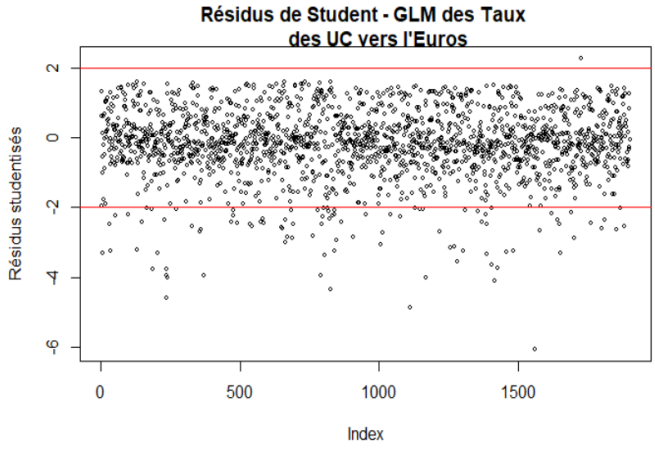

FIGURE 4.10 – Résidus de Student : Arbitrages des fonds UC vers le fonds euro

2 L'arbre CART : Arbitrages des fonds UC vers le fonds euro

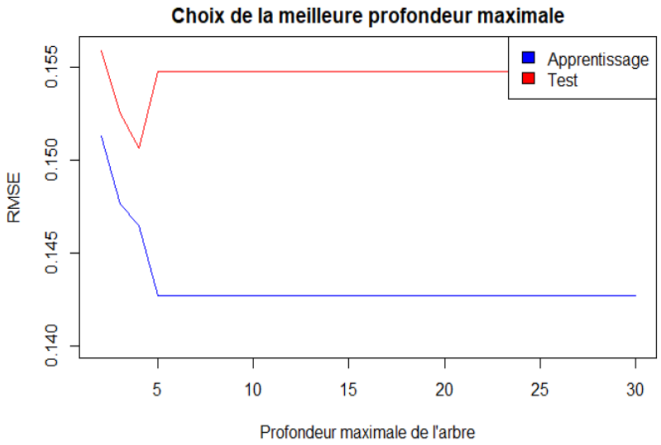

FIGURE 4.11 – Choix de la profondeur maximale de l'arbre : Arbitrages des fonds UC vers le fonds euro

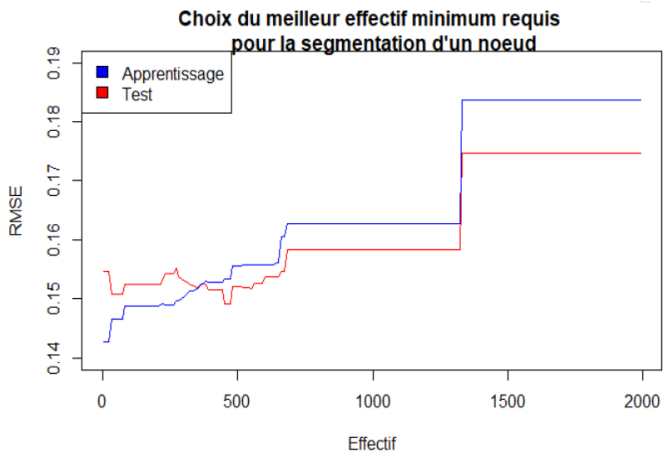

FIGURE 4.12 – Choix de l'effectif minimum par nœud : Arbitrages des fonds UC vers le fonds euro

3 La forêt aléatoire : Arbitrages des fonds UC vers le fonds euro

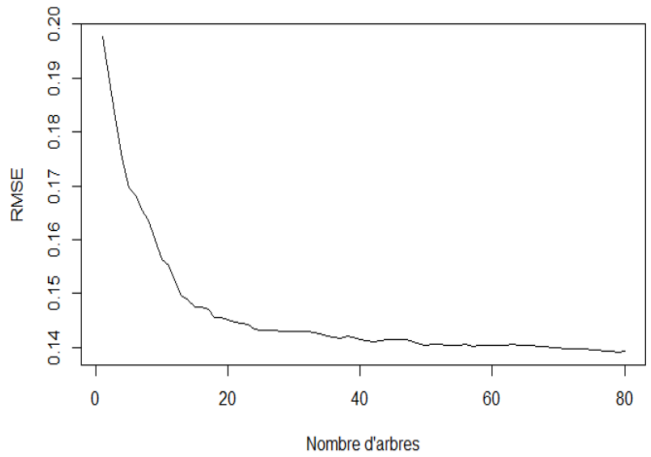

FIGURE 4.13 – Choix du nombres d'arbres de la forêt : Arbitrages des fonds UC vers le fonds euro

4 Choix du meilleur modèle : Arbitrages des fonds UC vers le fonds euro

Modèle	RMSE	MSE	MAE	MAPE
Modèle moyen	0,174	0,030	0,135	2,788
GLM	0,1475	0,022	0,109	3,186
Arbre CART	0,152	0,023	0,109	2,449
Forêt aléatoire	0,137	0,019	0,097	2,337

TABLE 4.1 – Différents indicateurs calculés pour les modèles

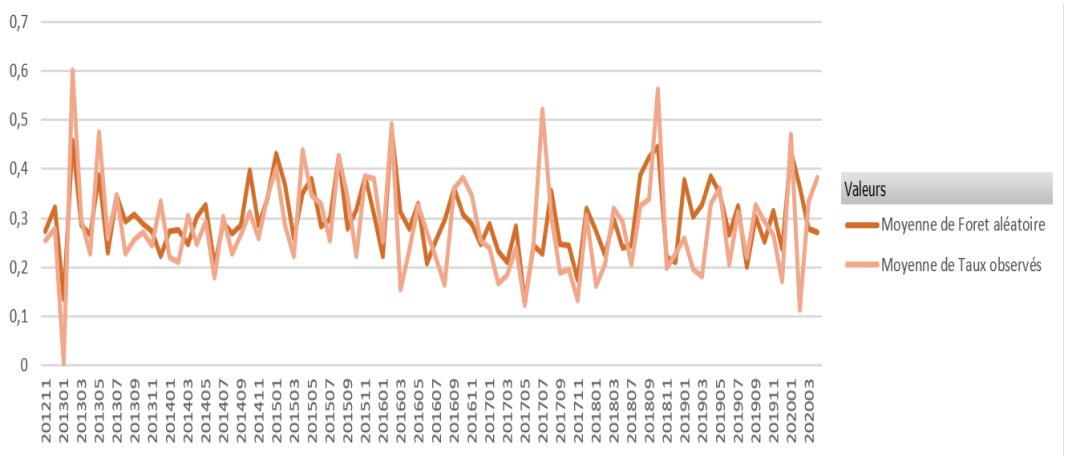

FIGURE 4.14 – Prédictions moyennes de la forêt aléatoire : Arbitrages des fonds UC vers le fonds euro

4. Bibliographie

[1] Leo Breiman. *Bagging predictors.* 1996.

[2] Leo Breiman. *Random Forests.* 2001.

[3] Cabinet d'actuariat Actuelia. Rapports narratifs, rsr et sfcr.

[4] Yadolah Dodge. *Analyse de régression appliquée.* 1999.

[5] Leo Breiman et al. *Classification and regression trees.* 1984.

[6] J.-M. Azais et J.-M.Bardet. *Le modèle linéaire par l'exemple.* 2005.

[7] Robin Genuer et Jean-Michel Poggi. *Arbres CART et Forêts aléatoires, Importance et sélection de variables.* 2017.

[8] D. B. Rubin et R. J. Little. Statistical analysis with missing data, 2002.

[9] Pierre-Louis GONZALEZ. L'analyse en composantes principales.

[10] Simone Manganelli and Robert Engle. *Value at Risk Models in Finance.* 2001.

[11] Frédéric PLANCHET Marc JUILLARD. Pilier 2 : vers le pilotage d'un profil de risques. 2010.

[12] Site de l'ACPR (Autorité de contrôle prudentiel de résolution). Solvabilité 2.

[13] Stéphane Tufféry. *Data Mining Et Statistique Décisionnelle - L'intelligence Des Données.* 2012.