IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management pour la validation du cursus à la Formation d'Actuaire de l'Institut du Risk Management et l'admission à l'Institut des actuaires le

Par : Jabri Othman

Titre : Application des méthodes de Data Mining à l'agrégation du passif en prévoyance et santé

Confidentialité : ☑ NON ☐ OUI (Durée : ☐ 1an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise : Malakoff Humanis

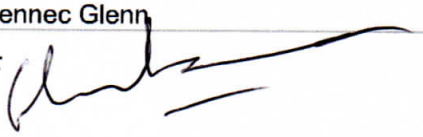
Nom :

Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : Favennec Glenn

Signature :



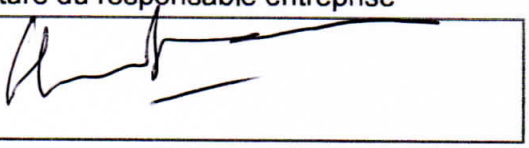
Invité :

Nom :

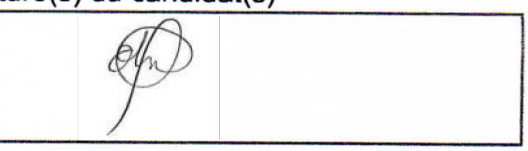
Signature :

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(s) du candidat(s)



Résumé — Ce mémoire présente une nouvelle méthode d'agrégation de passifs techniques d'assurance appliquée à un portefeuille de contrats prévoyance et santé.

Sur le plan méthodologique, l'approche proposée se distingue des techniques d'agrégation usuellement mises en œuvre dans les entreprises pour deux raisons principales.

D'une part pour construire l'outil d'agrégation, les critères de similarité entre risques ne s'appliquent pas aux caractéristiques intrinsèques des individus modélisés mais aux cash-flows futurs qui leur sont associés. Cela permet d'agréger des passifs relatifs à des individus ayant des caractéristiques intrinsèques différentes (hommes et femmes d'âges différents par exemple). En particulier et contrairement aux techniques usuelles, la performance de l'approche proposée croît avec le nombre de paramètres associés aux individus et donc avec le caractère élaboré de la modélisation.

D'autre part, en exploitant les possibilités offertes par les outils d'apprentissage automatisé, l'approche proposée permet de construire un outil insensible au vieillissement du portefeuille : il peut donc être utilisé à différentes dates d'arrêté, par exemple dans le cadre du processus de production des arrêtés trimestriels.

Les résultats numériques présentés dans le mémoire font apparaître une performance satisfaisante : l'impact est très faible sur les indicateurs financiers au regard de la diminution substantielle de la taille du passif projeté.

Mots clés — Prévoyance et santé, Modélisation actuarielle, Solvabilité II, Classification, Apprentissage automatisé.

Abstract — In this paper, we introduce a new method of aggregating technical insurance liabilities applied to income protection and health line of business.

The proposed approach differs from the common aggregation techniques for two main reasons.

On the one hand, the criteria of similarity between risks do not rely on the characteristics of the policyholders but on their future cashflows : this aspect empowers the method and allows to aggregate liabilities associated to policyholders with widely different characteristics (for instance, we can have inside the same cluster men and women with different ages). Consequently and unlike common techniques, the performance of the proposed approach increases with the number of modelling parameters taken into account. In other words, the efficiency of the method is more important when the modeling approach is more accurate.

On the other hand, based on machine learning algorithms, it was possible to make the proposed methodology stable against the aging of the portfolio, which allows to apply it sustainably for quarterly financial reporting.

The results obtained show a satisfying performance and a very low impact on the different financial indicators.

Keywords — Income protection and health, Actuarial modelling, Solvency II, Clustering, Machine learning.

Remerciements

Mes remerciements s'adressent à l'ensemble des intervenants de la formation d'Actuaire de l'Institut du Risk Management qui nous ont enseigné avec enthousiasme et implication toutes les connaissances et les compétences nécessaires à la réalisation de la présente étude.

Je tiens également à remercier M. Glenn Favennec, mon directeur de mémoire pour son aide précieuse, pour son encadrement tout au long des travaux ainsi que pour la pertinence de ses conseils.

Mes remerciements se tournent aussi vers tous les membres de la direction de la Solvabilité de Malakoff Humanis pour leur accompagnement pendant la réalisation des travaux dans une ambiance conviviale, chaleureuse et motivante.

Très bonne lecture.

Table des matières

Liste des sigles et acronymes

EIOPA	*European Insurance and Occupational Pensions Authority*
MPA	*Model Point d'actif*
MPP	*Model Point de passif*
BEL	*Best Estimate Liabilities*
FDB	*Future Discretionary Benefits*
ORSA	*Own Risk And Solvency Assessment*
QRT	*Quantitative Reporting Templates*
IFRS	*International Financial Reporting Standards*
S/P	*Sinistres / Primes (Loss ratio)*
ALM	*Asset Liability Management*
MCEV	*Market Consistent Economic Value*
OPCVM	*Organisme de Placement Collectif en Valeurs Mobilières*
BCAC	*Bureau Commun d'Assurances Collectives*
MGDC	*Maintien de la garantie décès*
IT	*Incapacité temporaire*
IP	*Invalidité permanente*
RE	*Rente d'éducation*
RC	*Rente de conjoint*
CoC	*Cost Of Capital*
RM	*Risk Margin*

Introduction

L'évaluation des provisions Best Estimate et du bilan économique Solvabilité II requiert la projection des flux de trésorerie liés à chaque contrat figurant au passif d'une compagnie d'assurance avec des outils de modélisation actuarielle. Compte-tenu du caractère complexe que présentent certaines garanties offertes aux assurés ainsi que du nombre très important de contrats détenus par les grands groupes d'assurance, l'approche de projection en tête par tête pour les dossiers de sinistre pose plusieurs problématiques de temps de calcul, de robustesse des plate-formes informatiques et de complexité d'analyse.

Le règlement délégué de la Commission Européenne [C.E 2014] ainsi que les orientations [EIOPA 2014] relatives au calcul de la meilleure estimation des provisions précisent que les assureurs peuvent recourir à des méthodes d'agrégation des passifs à condition que ces derniers présentent une nature de risque identique et que l'agrégation n'ait pas d'impact significatif sur les résultats. De surcroît, des délais de plus en plus courts sont exigés pour les arrêtés trimestriels et annuels. Par conséquent, l'ensemble des étapes du processus de production du bilan économique, notamment celui de l'agrégation du passif est soumis à de fortes contraintes d'optimisation.

Les bases de données massives de contrats à projeter ouvrent le champ des possibles en termes de méthodes d'exploration de données pouvant être utilisées pour l'agrégation du passif. Néanmoins, la modélisation est différente pour chaque type de contrat et chaque portefeuille possède des caractéristiques spécifiques, il convient donc d'adapter le choix des méthodes en fonction de la nature du portefeuille. En outre, la méthodologie d'agrégation pouvant faire l'objet d'un contrôle par le régulateur, son impact sur le bilan économique Solvabilité II doit être maîtrisé.

De ce fait, quelles méthodes statistiques peuvent être mises en place pour optimiser l'agrégation du passif ? Quel est l'impact sur les différents indicateurs actuariels et sur le bilan économique ? Finalement, comment ces méthodes peuvent-elles s'inscrire dans le cadre des processus trimestriels de production financière ?

Dans le cadre de ce mémoire, nous allons nous intéresser à un portefeuille de contrats de prévoyance et santé, offrant des garanties en cas de décès, d'arrêt de travail et de remboursement de frais de soins.

Chapitre 1

Cadre général, objectifs et démarche actuarielle

1.1 Mise en contexte

Sous la directive Solvabilité II, le bilan économique d'une compagnie d'assurance se présente de la manière suivante :

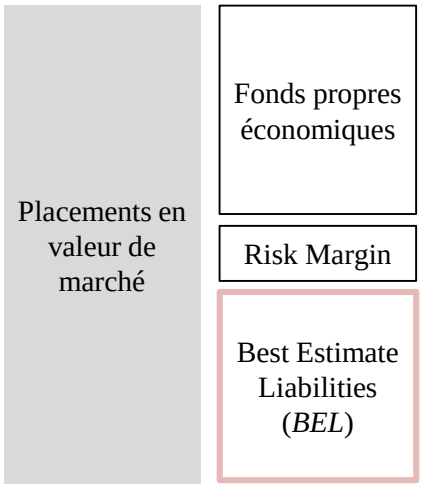


Figure 1.1. Bilan Solvabilité II simplifié

Les travaux de ce mémoire s'inscrivent dans le cadre du calcul du Best Estimate Liabilities (BEL). Cela correspond à la meilleure estimation des engagements techniques de l'assureur. Il s'obtient en actualisant, au taux sans risque, la différence entre les flux probables sortant et les flux probables entrants, sur un horizon de projection T, fixé en fonction de la duration du portefeuille.

$$BEL = \sum_{t=1}^{T} \frac{cashflow_{out}(t) - cashflow_{in}(t)}{(1+r_t)^t} \tag{1.1}$$

Les $cashflow_{in}$ comprennent l'ensemble des flux de trésorerie positifs tels que les primes et les produits financiers. Les $cashflow_{out}$ correspondent aux flux négatifs dûs par l'assureur, par exemple les prestations à payer en cas de sinistre et ses frais de fonctionnement.

Les cashflows futurs sont évalués à l'aide d'un outil de projection actuarielle, et ce à partir de différentes données en entrée ou inputs :

- **Les Model Points :** Les Model Points Actifs (MPA) correspondent aux titres financiers détenus par l'assureur, les Model Points Passif (MPP) correspondent aux contrats souscrits par la compagnie ;
- **Les hypothèses financières :** elles correspondent aux données économiques à prendre en compte telles que la courbe des taux d'intérêts ;
- **Les hypothèses techniques :** hypothèses applicables au passif, par exemple les lois de mortalité des assurés ou de rachat des contrats.

Ces éléments alimentent le modèle de projection, en d'autres termes les algorithmes et les règles de calcul permettant de modéliser les flux futurs probables ainsi que l'interaction Actif-Passif. Schématiquement, le processus de production du bilan Solvabilité II d'une compagnie d'assurance peut se présenter de la manière suivante :

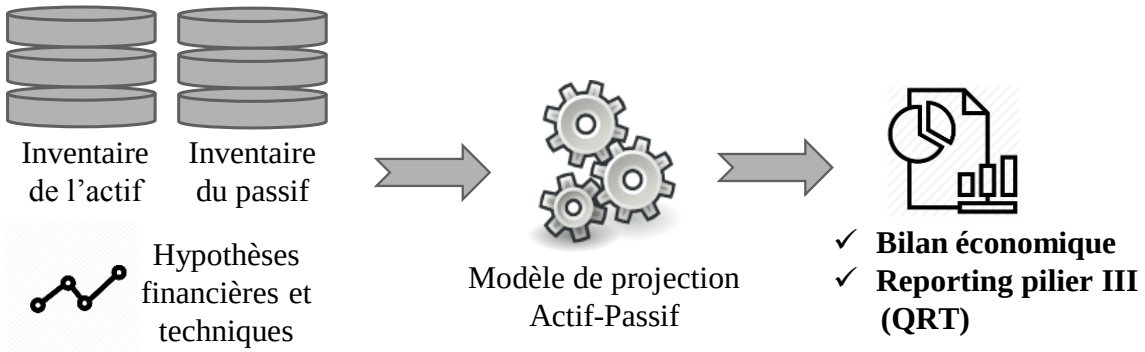


Schéma 1.1. Processus de production du Reporting Solvabilité II

Le fichier de Model Points de passif doit contenir des informations relatives à tous assurés. Néanmoins, les portefeuilles des grands groupes d'assurance contiennent un nombre très important d'assurés, ce qui rend l'approche de projection en tête par tête très contraignante du point de vue de la capacité des ressources informatiques nécessaires à l'exécution des modèles (notamment en termes de temps de calcul et de consommation en mémoire).

Si certains assureurs disposent aujourd'hui de ressources informatiques leur permettant de faire face aux problèmes engendrés par la dimension du passif à projeter, il n'en reste pas moins que le coût financier généré par ces ressources est non négligeable.

Le règlement délégué de la Commission Européenne dispose que les assureurs peuvent avoir recours au regroupement des contrats à projeter, sous contrainte de perte minimale de la précision dans les calculs : cela motive donc l'intérêt porté à la problématique de l'agrégation du passif.

Dans ce mémoire, nous allons nous intéresser à l'agrégation d'un portefeuille de prévoyance et santé collectives dans le cadre des calculs des indicateurs Solvabilité II : *Best Estimate Liabilities (BEL) et Solvency Capital Requirement (SCR)*

Au-delà de l'aspect purement financier, étudier la problématique d'agrégation du passif présente plusieurs intérêts :

- D'abord, cela permet d'alléger le processus de production trimestriel Solvabilité II et donc de consacrer plus de temps à l'analyse des résultats ;
- La projection d'un nombre restreint de Model Points est une approche plus simple d'un point de vue opérationnel compte-tenu de la taille des fichiers manipulés quand la projection s'effectue en tête par tête ;
- Plusieurs applications sont rendues possibles par l'agrégation du passif, notamment la construction de profils d'assurés représentatifs pour réaliser les calculs dans le cadre de l'ORSA et de traiter plus facilement d'autres sujets tels que les études ALM avec un passif léger et facilement exploitable.

Par ailleurs, les résultats de cette étude peuvent être exploités plus largement dès qu'il s'agit de projeter un nombre très important de contrats, par exemple pour d'autres normes de Reporting et de communication financière comme les IFRS17 ou MCEV.

1.2 Revue bibliographique

Depuis le lancement de l'exercice préparatoire pour la mise en place de la directive Solvabilité II, la question de l'agrégation du passif pour les besoins de calcul du Best Estimate a été évoquée chez la majorité des acteurs de l'assurance. Cette problématique a été également traitée dans de nombreux mémoires d'actuariat.

Avant de détailler les ressources bibliographiques en lien avec cette problématique, il convient d'abord de mentionner que la méthode d'agrégation à utiliser dépend fortement de l'approche de modélisation des contrats. La modélisation des contrats de prévoyance et santé sera détaillée dans les chapitres suivants, cependant, dans le but de bien comprendre

la bibliographie, notons qu'il existe aujourd'hui deux approches courantes de modélisation que nous pouvons résumer brièvement de la manière suivante :

- **Modélisation en S/P :** l'approche S/P consiste à suivre en moyenne l'évolution des grandeurs projetées (provisions, prestations, frais, commissions, etc.), sans qu'un décompte des effectifs des assurés ne soit effectué ;
- **Modélisation Markovienne :** elle consiste à simuler, pour un état initial donné (par exemple l'invalidité), l'ensemble des transitions probables d'une tête assurée vers un autre état. A cet effet, le modèle de projection permet de projeter l'effectif de chaque état et de calculer les flux futurs probables.

Pour la modélisation en S/P, le problème d'agrégation des contrats revient à identifier la maille de calcul conduisant à des hypothèses à la fois peu volatiles et reflétant au mieux les spécificités de chaque groupe de contrats identifié. Le mémoire [Tann 2011] (cité dans les références bibliographiques) fournit une méthode d'optimisation du niveau de segmentation d'un portefeuille de prévoyance individuelle, reposant sur la construction des hypothèses. En effet, le niveau de segmentation optimal correspond à la maille adéquate de définition des hypothèses.

Pour ce qui est de la modélisation Markovienne, l'agrégation des dossiers de sinistres repose généralement sur les caractéristiques intrinsèques des assurés : par exemple, l'âge et l'ancienneté dans l'état de sinistre pour les garanties arrêt de travail (cf. 2.1). Le principe de base consiste à trouver des groupes homogènes d'assurés en utilisant ces variables d'intérêt. Ceci dit, plusieurs variantes méthodologiques reposant sur ce même principe ont été proposées dans la littérature : elles se distinguent par le choix des méthodes statistiques, les critères de validation de la segmentation obtenue ainsi que par les indicateurs permettant de mesurer l'impact de l'agrégation.

Si les travaux de constitution de groupes homogènes sur un portefeuille de prévoyance de [Le Corre 2012] reposent également sur les caractéristiques intrinsèques des assurés, il n'en reste pas moins qu'ils fournissent une première variante méthodologique. Après une étude détaillée sur le portefeuille des assurés (statistique descriptive, distribution des variables d'intérêts), la construction de groupes homogènes d'assurés est réalisée par la méthode de la classification ascendante hiérarchique. En outre, l'utilisation des méthodes de classification supervisée (réseaux de neurones, méthodes SVM, régression logistique) est évoquée, essentiellement dans le but de valider la segmentation obtenue. Le mémoire [Couloumy 2013] présente une méthode d'agrégation d'un portefeuille reposant pareillement sur les caractéristiques intrinsèques des assurés, toutefois, on y trouve une introduction de l'analyse en composantes principales afin de réduire le nombre de variables entrant dans la classification. Le mémoire fournit également un large panel de stratégies d'agrégations, incluant les impacts sur le Best Estimate de chaque stratégie.

Les travaux de [Goffard 2015] introduisent une nouvelle approche d'agrégation du pas-

sif appliquée à un portefeuille d'épargne individuelle. Contrairement aux autres méthodes d'agrégation s'appliquant sur les caractéristiques intrinsèques des assurés, la procédure consiste à appliquer les méthodes de classification non-supervisée (CAH pour l'initialisation, K-means pour l'agrégation puis construction de profils moyens représentatifs) sur les probabilités de sortie des assurés au cours de la projection (suite à un rachat, un décès ou une arrivée à échéance).

1.3 Objectifs du mémoire

Nous avons vu que les méthodes d'agrégation du passif proposées dans la littérature s'appuient généralement sur les caractéristiques des assurés (agrégation sur la base du sexe, de l'âge ou d'autres paramètres caractérisant les assurés). D'autre part, la question de la pérennité de la méthode d'agrégation reste globalement à approfondir.

Compte-tenu de l'état actuel de la recherche, l'objectif de notre étude est de développer une nouvelle approche d'agrégation du passif en prévoyance et santé : une approche efficace, stable dans le temps et s'appuyant sur des hypothèses invariantes. D'autre part, le but est de tirer profit de l'ensemble des idées intéressantes que l'on a pu voir dans les travaux précédents.

D'abord, plutôt que de construire les groupes homogènes en utilisant les caractéristiques intrinsèques des assurés (âge, ancienneté, sexe etc.), nous allons fonder la construction des groupes sur les cashflows qui leurs sont associés, tels que les prestations futures probables ou les provisions mathématiques projetées : le choix de la grandeur adéquate à utiliser pour l'agrégation sera fait en fonction de la nature et des paramètres impactant le risque étudié.

Cette démarche nous permettra de réaliser des agrégations plus intéressantes que celles obtenues sur la base des caractéristiques intrinsèques des assurés (un exemple concret sera fourni ultérieurement dans la section 5.1).

Ensuite, un critère de validation applicable aux flux agrégés sera retenu pour vérifier la robustesse de notre méthode. Bien entendu, les indicateurs à contrôler in fine seront le BEL, le SCR ainsi que le ratio de couverture Solvabilité II.

Par ailleurs, étant données les contraintes de Reporting trimestriel, la procédure d'agrégation qui sera construite doit s'inscrire dans la durée, c'est-à-dire qu'elle doit offrir la possibilité d'agréger le portefeuille de contrats avec une fréquence régulière, malgré son évolution et son vieillissement au cours du temps. Deux aspects seront donc respectés pour répondre à cette exigence :

1– Appliquer l'agrégation sur des grandeurs issues d'hypothèses généralement stables dans le temps ou du moins peu volatiles (par exemple, les tables de mortalité réglementaires) ;

2– Disposer d'un algorithme d'agrégation entraîné sur une base de données adéquate, et ce grâce aux méthodes de classification supervisée. L'objectif étant de pouvoir l'appliquer sur les portefeuilles mis à jour à chaque arrêté prudentiel.

1.4 Démarche retenue

A ce stade, nous avons défini la problématique et les objectifs du mémoire en considérant les études existantes.
Pour atteindre notre objectif de construction d'une approche d'agrégation du passif, nous allons dans un premier temps présenter le portefeuille étudié. Plus précisement, nous allons nous intéresser à l'analyse statistique des données à disposition, tout en veillant à effectuer des contrôles de cohérence. Dans cette partie, il sera également question d'avoir une première idée sur la distribution des variables d'intérêt ainsi que leur impact sur la projection des flux futurs.

Cette projection se fait à l'aide d'un outil de modélisation actuarielle. L'étape suivante consistera donc à présenter le modèle de projection et les hypothèses retenues. En particulier, nous allons nous pencher sur les méthodes de modélisation des garanties prévoyance et santé ainsi que sur le détail des flux futurs de trésorerie qui nous permettront d'une part d'agréger les contrats, et d'autre part de calculer les indicateurs Solvabilité II de l'entité que nous étudions.

Par ailleurs, étant donné que notre approche d'agrégation du passif repose notamment sur les méthodes statistiques d'exploration de données *(Data Mining)*, le chapitre suivant du mémoire consiste à présenter la théorie de la classification supervisée et non-supervisée.

Dès que l'ensemble des éléments permettant de construire notre approche d'agrégation sont définis, nous l'appliquons sur le portefeuille après avoir exposé en détail le principe de la méthode ainsi que les étapes mises en place lors de la déclinaison opérationnelle. Nous tenons à préciser à ce niveau que plusieurs méthodes ont été testées pour optimiser l'agrégation, notamment des variantes aux méthodes statistiques de classification.

Finalement, la dernière étape de notre étude consiste à analyser l'impact de l'agrégation sur le ratio de couverture. Nous allons donc avoir recours au calcul des fonds propres économiques ainsi que des sensibilités techniques et financières du portefeuille agrégé, sur la base des différents chocs proposés par la formule standard Solvabilité II.

Chapitre 2

Portefeuille et données

2.1 Présentation des garanties étudiées

Nous nous intéressons à un portefeuille d'assurance de personnes et particulièrement de prévoyance et santé complémentaires. Il s'agit du régime de protection sociale complétant le régime obligatoire de la sécurité sociale. Ce secteur présente une grande diversité d'acteurs (sociétés d'assurances, institutions de prévoyance et mutuelles) pour un montant total de 34,2 milliards d'euros de cotisations en 2017 (Source : Fédération Française de l'Assurance). Du point de vue réglementaire, ce régime est soumis au code de la sécurité sociale, à la loi Evin et à l'Accord National Interprofessionnel.
Les principales garanties couvertes par les contrats étudiés sont récapitulées ci-dessous :

Risque	Garantie	Nature de l'engagement
Prévoyance - Arrêt de travail	**Incapacité temporaire (IT)**	Couverture de l'arrêt de travail (indemnité journalière) en cas d'incapacité temporaire
	Invalidité permanente (IP)	Versement d'une rente mensuelle en cas d'invalidité permanente
Santé	**Maladie (MAL)**	Remboursement des frais de soins de santé
Prévoyance - Décès	**Décès (DC)**	Temporaire décès (versement d'un capital) – contrats annuels
	Maintien de la garantie décès (MGDC)	Maintien des garanties décès suite à un arrêt de travail
	Rente d'éducation (RE)	Versement d'une rente éducation en cas de décès du parent
	Rente de conjoint (RC)	Versement d'une rente de conjoint en cas de décès du conjoint

Table 2.1. Risques et garanties offertes aux assurés au sein du portefeuille étudié

2.2 Composition du portefeuille

Dans cette partie, nous ferons une brève description de l'actif et du passif avant de fournir le bilan comptable associé au portefeuille étudié.
Par souci de confidentialité, **les données affichées ne correspondent pas aux données réelles de l'entité étudiée : il s'agit d'un portefeuille fictif, reflétant les caractéristiques classiques des contrats de prévoyance et santé**. L'unité des montants présentés n'est pas précisée car elle n'a pas d'impact sur les conclusions de l'étude.

Composition de l'actif

Les placements financiers sont en représentation des engagements d'assurance ainsi que des capitaux propres.

Les actifs détenus par l'entité sont regroupés selon deux catégories dans la base des Model Points d'actifs. D'une part, la catégorie *Equity* couvre :
–Les actions ou les parts de capitaux propres d'une entreprise externe.
–Les actifs immobiliers constitués d'immeubles d'exploitations et d'investissement.
–Les OPCVM (Organismes de Placement Collectif en Valeurs Mobilières) dont les fonds investis sont placés en valeurs mobilières ou autre instruments financiers.
D'autre part, la catégorie *Bond* regroupe les obligations à taux fixe, les obligations à taux variable et les obligations à taux indexé sur l'inflation. Il s'agit donc des instruments de dette assortie d'un taux d'intérêt donnant droit à la perception de coupons à une fréquence et une échéance déterminées.

Le tableau ci-après donne une synthèse des proportions et des montants des actifs en vision comptable. Les valeurs de marché des placements sont également présentées dans l'optique de construire le bilan économique Solvabilité II.

	Valeur comptable totale 3 020 653	**Valeur de marché totale 3 285 708**
Obligations	73%	75%
Actions	8%	7%
OPCVM	9%	7%
Immobilier	10%	11%

Table 2.2. Composition de l'actif

Composition du passif

Le passif comptable est composé des capitaux propres et des provisions d'inventaire *(French GAAP)*.
Une provision technique est calculée pour chaque garantie. Elle correspond, selon les garanties, à une provision mathématique ou à une provision pour sinistres à payer.
Par ailleurs, les contrats collectifs étudiés prévoient une clause de participation aux bénéfices, deux provisions sont constituées à ce titre : la provision d'égalisation et la réserve générale.
Le tableau suivant donne la définition de chaque provision, son montant ainsi que son périmètre de calcul :

		Arrêt de travail	Décès	Santé
Provision Mathématique (PM)	Provision constituée pour le paiement des prestations futures probables des individus en portefeuille à la date d'évaluation	1 071 824	430 656	
Provision pour sinistres à payer (PSAP)	RBNS *(reported but not settled)* : Provision constituée pour le paiement des prestations au titre des sinistres connus déclarés à la date d'évaluation IBNR *(incurred but not reported)* : Provision constituée pour le paiement des prestations au titre des sinistres non connus et non déclarés à la date d'évaluation	186 981	222 624	245 847
Provision d'égalisation (PE)	Provision constituée pour pallier les fluctuations adverses de sinistralité future sur la prévoyance	18 473		
Réserve générale (RG)	Provision constituée pour pallier les fluctuations adverses de sinistralité future sur la santé, également alimentée des excédents fiscaux de la PE	44 283		

Table 2.3. Composition du passif

Bilan French GAAP

Nous avons présenté l'ensemble des éléments permettant de construire notre bilan comptable en normes locales *(French GAAP)*. A ce stade, les placements sont comptabilisés en

valeur nette comptable et les provisions selon les règles Solvabilité I. Les capitaux propres sont quant à eux obtenus par différence (Placements - Provisions).

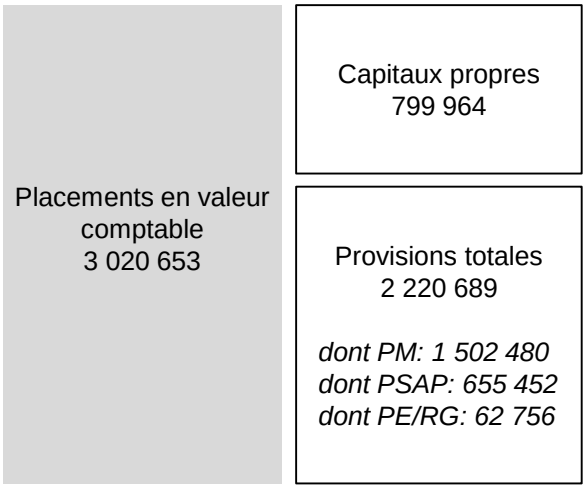


Figure 2.1. Bilan French GAAP

2.3 Model Points de passif

L'objectif du mémoire étant de construire une méthode d'agrégation du passif, une description détaillée du fichier d'inventaire du passif est fournie dans cette section, notamment les informations qu'il contient ainsi que sa dimension.

Tout d'abord, pour les besoins de modélisation et de Reporting, les informations contenues dans le fichier de passif sont présentées par garantie et par segment Solvabilité II.
Le fichier comporte 66 212 lignes. En plus de la segmentation par garantie, chaque ligne appartient à l'une des deux catégories suivantes :

- **Portefeuille en stock :** données relatives au stock de contrats, elles regroupent les assurés et les contrats existants en portefeuille à la date d'évaluation N. Les provisions (PM et PSAP) liées à ces contrats sont celles reportées dans le bilan *French GAAP* défini précédemment ;
- **Primes futures :** données relatives aux cotisations futures des assurés. En effet, conformément à la notion Solvabilité II de la frontière des contrats, les primes du budget l'année N+1 sont intégrées dans le calcul des engagements de l'assureur en prévoyance et santé.

De manière générale, les données à disposition sont des données en tête par tête, c'est-à-dire que chaque ligne du fichier d'inventaire correspond à un seul assuré couvert par l'entité, à l'exception des :

– Model Points relatifs aux primes futures ;
– Garanties faisant appel uniquement aux provisions pour sinistres à payer et non aux provisions mathématiques.

Pour ces deux dernières catégories, chaque ligne correspond à un profil moyen d'assurés. La dimension du fichier – le nombre de lignes par garantie et par catégorie – est présentée dans le tableau suivant :

Catégorie	Garantie	Nombre de MPP	% nombre total	Nature du MPP
Stock de contrats	**Invalidité**	19 443	*29,4%*	Tête par tête
	Incapacité	16 419	*24,8%*	
	MGDC	19 706	*29,8%*	
	Rente de conjoint	3 916	*5,9%*	
	Rente éducation	4 067	*6,1%*	
	Temporaire décès	1 101	*1,7%*	Profil moyen
	Santé	231	*0,3%*	
Primes futures	**Toutes garanties**	1 329	*2,0%*	
Total		**66 212**		

Table 2.4. Dimension du fichier d'inventaire du passif

Les colonnes du fichier fournissent les paramètres nécessaires pour la modélisation et pour la projection des cashflows futurs. Ces paramètres dépendent principalement de la garantie et de la nature de la ligne en question :
– Pour la modélisation en tête par tête, les colonnes contiennent les caractéristiques intrinsèques de l'assuré, par exemple, pour les rentes de conjoint, nous avons besoin entres autres de l'âge du conjoint et du montant de la rente
– Pour la modélisation en profil moyen, les colonnes contiennent les références aux hypothèses à utiliser pour la projection, par exemple le type de la cadence de règlement des PSAP ou l'hypothèse de sinistralité (S/P) à appliquer au montant des primes

L'agrégation du passif portera exclusivement sur les lignes modélisées en tête par tête, car :
–D'une part, la maille de projection optimale pour les profils moyens correspond simplement à celle de la construction des hypothèses (cf. [Tann 2011])
–D'autre part, le nombre de lignes modélisées en profil moyen n'est pas matériel ($\leq 5\%$)

2.4 Statistique descriptive et contrôles de cohérence

Maintenant que nous avons une idée globale sur la composition du fichier à disposition, le but de cette section est d'analyser les données relatives à chaque garantie : à savoir établir une statistique descriptive des variables d'intérêt et comprendre leur distribution.

Une remarque importante à signaler à ce stade : les variables qui seront analysées pour chaque garantie ne correspondent pas, dans l'absolu, aux seuls paramètres impactant la garantie dans un cadre plus général. En effet, l'objectif est de concentrer l'analyse sur les paramètres impactant la garantie, compte-tenu des hypothèses et de l'approche de modélisation retenues dans le cadre de notre étude.

Invalidité

Comme souligné précédemment, cette garantie représente 29,4% des Model Points disponibles dans le fichier de passif à agréger. Il s'agit des assurés en stock dont l'état initial à la date d'évaluation correspond à l'invalidité permanente. Pour cette garantie, l'assureur prévoit de verser un forfait mensuel jusqu'à l'échéance prévue par le contrat. Cette échéance est égale à 62 ans pour l'ensemble des assurés en portefeuille.

A ce titre, une provision mathématique est constituée par l'assureur. De plus, une provision pour sinistre à payer est prévue pour cette garantie, elle permet de faire face aux sinistres non encore déclarés et aux retards de paiement.

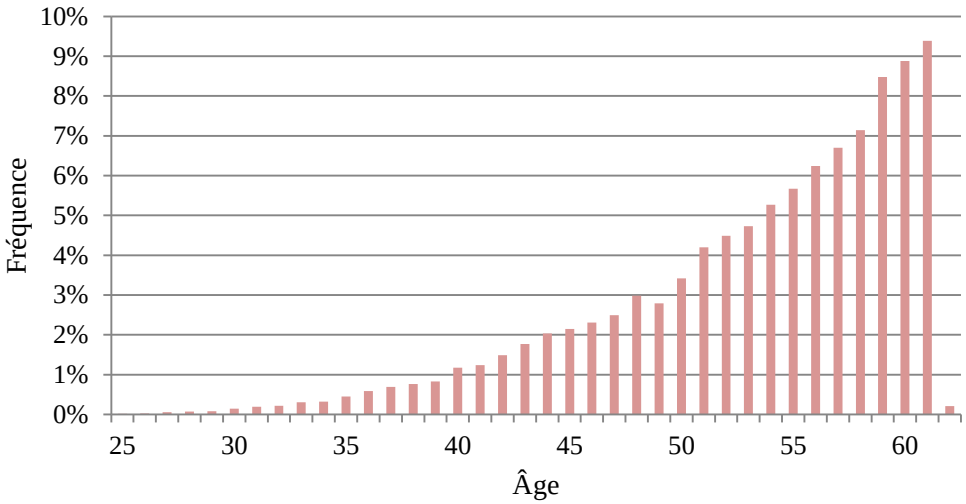
En termes de matérialité, les provisions constituées pour cette garantie sont données par le tableau ci-dessous. Pour rappel, le montant des provisions totales du portefeuille, toutes garanties confondues, est de 2 220 689.

	Provisions totales	PM	PSAP
Invalidité	807 143	718 988	88 154

Table 2.5. Provisions Solvabilité I - Invalidité

En rapportant la provision mathématique au montant total des rentes, nous trouvons un indicateur homogène à une duration moyenne dans l'état d'invalidité. Sur le portefeuille étudié, les rentes d'invalidité durent 6 ans en moyenne.

Nous allons à présent nous pencher sur les propriétés statistiques des données des assurés en invalidité, pour ce faire nous nous intéressons à deux variables. La première étant l'âge des assurés, sa distribution est donnée par le graphe ci-après.



Graphique 2.1. Distribution de l'âge des assurés en invalidité

Force est de constater que l'effectif des assurés en invalidité augmente avec l'âge, à tel point que plus de 75% des effectifs ont un âge supérieur à 50 ans. L'âge moyen des assurés, pondéré par le montant de la rente servie, est quant à lui égal à 54 ans.

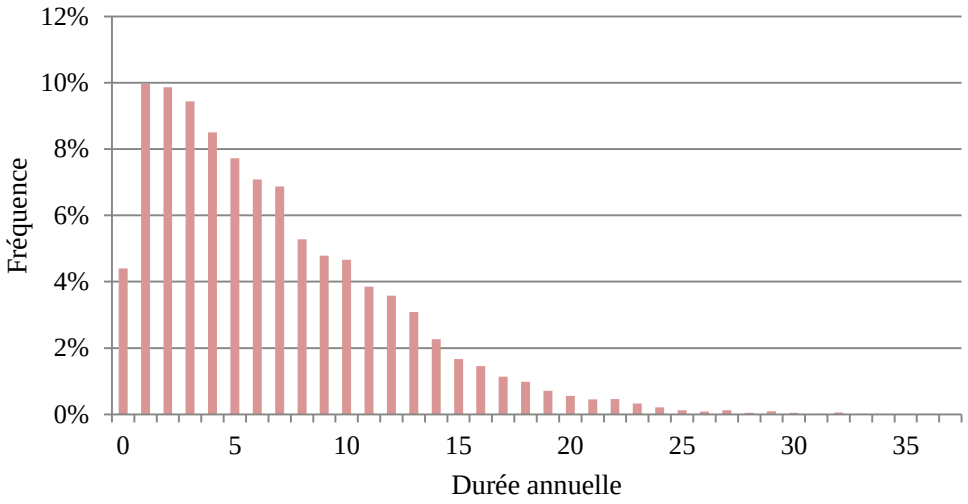
Un contrôle de cohérence de la provision mathématique a été réalisé à ce niveau, en vérifiant que la duration moyenne dans l'état d'invalidité à l'âge moyen est proche de l'indicateur obtenu en rapportant la provision mathématique au montant total de la rente.

La deuxième caractéristique qui nous intéresse est la date d'entrée en invalidité (cf. figure 2.2). Elle permet de calculer l'ancienneté de l'assuré dans l'état et par conséquent sa probabilité de maintien en invalidité.

La distribution obtenue pour l'ancienneté dans l'état d'invalidité a une allure inversée par rapport à celle de l'âge, en effet, les assurés entrant en invalidité étant généralement d'un âge relativement avancé, les effectifs d'ancienneté sont en conséquence centrés autour des faibles durées.

En combinant les deux variables (l'âge et l'ancienneté des assurés), nous pouvons calculer l'âge moyen d'entrée en invalidité, qui est de 47 ans sur le portefeuille étudié.

A titre indicatif, le pourcentage de femmes en invalidité dans le portefeuille étudié est légèrement supérieur à celui des hommes. Cela n'a pas d'impact dans le cadre de notre étude car les tables de maintien en invalidité qui seront utilisées ne font pas la distinction par sexe.



Graphique 2.2. Ancienneté des assurés en état d'invalidité

Incapacité

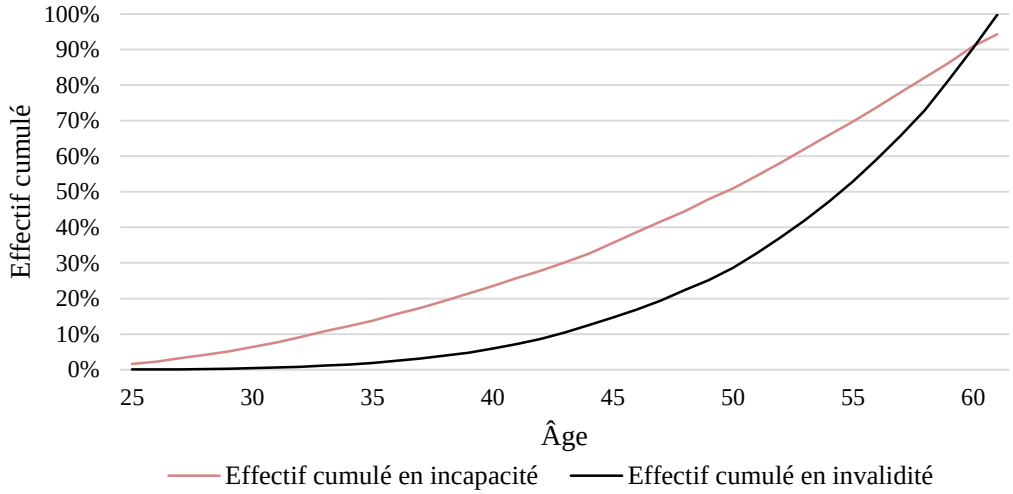
A l'instar de la garantie invalidité, deux types de provisions sont prévues pour payer les prestations aux assurés (une PM et une PSAP), leurs montants sont récapitulés dans le tableau qui suit :

	Provisions totales	PM	PSAP
Incapacité	447 178	348 886	98 293

Table 2.6. Provisions Solvabilité I - Incapacité

Nous nous intéressons pareillement à deux variables : l'âge des assurés et leur date d'entrée en incapacité. L'échéance est similaire à celle vue en invalidité, c'est-à-dire que les assurés sont couverts jusqu'à l'âge de 62 ans. Ceci dit, il existe des caractéristiques propres à chaque garantie au niveau des données des assurés.

Le nombre d'assurés en incapacité dans le portefeuille étudié est de 16 419 avec une moyenne d'âge égale à 48 ans. Nous remarquons que l'écart-type des âges en incapacité est plus élevé par rapport à celui de l'invalidité, autrement dit, les âges en incapacité sont moins centrés autour de leur valeur moyenne.
Un autre aspect permettant de comparer les données en incapacité et en invalidité est celui de la fonction de répartition de la distribution des âges. En effet, pour la garantie invalidité, nous avons vu que les effectifs augmentent considérablement à partir de 50 ans, en revanche nous observons une augmentation des effectifs plus uniforme en fonction de l'âge pour la garantie incapacité, comme le montre le graphe ci-dessous :



Graphique 2.3. Comparaison de la fonction de répartition des âges en invalidité/incapacité

Rappelons finalement que la couverture en incapacité est limitée dans le temps, en effet, les assurés ayant dépassé la durée de trois ans en incapacité sont classés directement en invalidité par le médecin conseil du régime de base. De manière générale, selon les tables du BCAC, seules 59% des personnes atteintes d'incapacité à l'âge de 48 sont maintenues dans l'état au bout du premier mois. En l'occurrence, sur le portefeuille étudié, plus de la moitié des assurés ont une ancienneté en incapacité inférieure à 12 mois, avec un âge moyen de 48 ans.

Maintien de la garantie décès

Cette garantie permet aux assurés en arrêt de travail de maintenir leur garantie en cas de décès tout au long de leur couverture au titre de l'incapacité ou de l'invalidité et même au delà de la durée prévue initialement par le contrat collectif (généralement d'une durée annuelle). En effet, la Loi Evin dispose que le maintien de la garantie décès est obligatoire.

Environ 29% des Model Points à notre disposition correspondent aux assurés couverts au titre de la garantie MGDC. Au niveau de la matérialité des provisions, le tableau 2.7 fourni une synthèse des provisions constituées au titre de cette garantie.

	Provisions totales	PM	PSAP
MGDC	157 366	136 293	21 074

Table 2.7. Provisions Solvabilité I - Maintien de la garantie décès

Pour simplifier, nous pouvons interpréter la garantie comme une temporaire décès à horizon 62 ans, chaque ligne de Model Point indique l'âge de l'assuré, le montant du capital

garanti ainsi que la date d'entrée en incapacité ou en invalidité. De ce fait, la population sous-jacente aux MPP de la garantie MGDC correspond d'une part à des assurés dont l'état initial est celui de l'incapacité, et d'autre part à des assurés en invalidité à la date d'évaluation.

Par conséquent, les mêmes propriétés statistiques que celles évoquées pour les garanties incapacité et invalidité sont vérifiées par les variables d'intérêt de la garantie MGDC. En d'autres termes, nous avons quasiment la même distribution pour l'âge et la date de sinistre. Le tableau ci-dessous permet de contrôler la cohérence des données en mettant en évidence la similarité entre la distribution des variables, à savoir des moyennes et des écarts-type très proches de ceux que l'on a pu observer pour les populations d'assurés en invalidité et en incapacité.

	Proportion	**Âge moyen**	**Ecart-type de l'âge**	**Ancienneté moyenne dans l'état**
MGDC – état initial: invalidité	42%	53	7,3	7
MGDC – état initial: incapacité	58%	48	10,1	1
MGDC – population totale		**50**	**9,4**	**3**

Table 2.8. Statistique descriptive pour le maintien de la garantie décès

Nous tenons à mentionner, avant de finir l'analyse statistique de cette garantie, que seule une partie de la population des assurés en incapacité/invalidité est incluse dans les MPP relatifs au maintien de la garantie décès. En effet, l'entité étudiée couvre, pour des contrats en acceptation, uniquement les risques d'arrêt de travail (incapacité/invalidité). D'autre part, l'entité peut avoir accepté par le passé d'assurer exclusivement les garanties décès sans pour autant porter les risques en arrêt de travail. Les risques portés par l'entité étudiée changent en fonction de la stratégie commerciale ainsi qu'en fonction du niveau de diversification souhaité, mais les caractéristiques des individus couverts sont stables dans le temps, si bien que la comparaison des distributions des âges et des anciennetés au sein des garanties MGDC, IT et IP a permis de confirmer cette stabilité.

Rente d'éducation

La garantie rente d'éducation est une garantie en cas de décès, elle a pour objectif de verser une rente, servant à financer les études des enfants dont le parent est décédé. Elle compte environ 4000 lignes de rentiers couverts par l'entité. Les montants des provisions

sont les suivants :

	Provisions totales	PM	PSAP
RE	72 991	57 158	15 833

Table 2.9. Provisions Solvabilité I - Rente d'éducation

Plusieurs paramètres entrent en jeu dans la projection des cashflows futurs relatifs à cette garantie :

- L'âge du bénéficiaire
  Cela permet de calculer la probabilité de survie ainsi que celle de la poursuite des études. L'âge moyen des rentiers est de 18 ans avec un écart-type de 5 ans. À noter que pour la table de poursuite des études, deux lois d'expérience seront utilisées, en fonction la catégorie socio-professionnelle du parent du bénéficiaire.
- Le sexe du bénéficiaire
  L'utilisation des tables générationnelles dans les projections Solvabilité II permet de tenir compte de l'impact du sexe sur le calcul des prestations futures. Sur le portefeuille étudié, il y a 52% de femmes et 48% d'hommes.
- Le terme, la fréquence et le type de la rente
  Les rentes servies sont à terme échu, avec une fréquence trimestrielle ou mensuelle. En outre, le terme de la rente d'éducation est fixé à 28 ans pour les contrats commercialisés en portefeuille.
- Le taux technique contractuel
  Le taux technique des rentes varie entre 0% et 5%, en fonction du taux moyen d'emprunt d'état au moment de la souscription du contrat. Ceci dit, la distribution des effectifs par taux technique montre que le niveau moyen du taux technique en rente d'éducation est de 1.8% sur le portefeuille étudié.

Rente de conjoint

Cette garantie prévoit le versement d'une rente au conjoint de l'assuré en cas de décès de ce dernier. Plus précisément, notre stock de contrats verse une rente trimestrielle ou mensuelle, viagère à terme échu au conjoint de l'assuré. Il y a 3 916 polices sinistrées en portefeuille, pour un montant de provisions détaillé dans le tableau 2.10.

Pour ce qui est de la distribution des âges des conjoints, la moyenne d'âge de la population étudiée est de 69 ans, avec un écart-type d'environ 12 ans, beaucoup plus large que celui observé en rente d'éducation. Par ailleurs, le coefficient d'asymétrie *(Skewness)* de la

distribution des âges en rente de conjoint est assez faible, plus précisément, il est de -18% en rente de conjoint et de -57% en rente d'éducation, ce qui signifie que la distribution des âges en rente d'éducation est très décalée à droite de la médiane. Ces aspects statistiques s'expliquent naturellement par l'échéance de la rente d'éducation servie, qui pour rappel, est fixée à 28 ans.

	Provisions totales	PM	PSAP
RC	282 512	241 749	40 764

Table 2.10. Provisions Solvabilité I - Rente de conjoint

De manière similaire à la rente d'éducation, un taux technique est défini pour chaque police. Il permet d'actualiser les prestations futures pour le calcul de la provision mathématique. Etant donné que le portefeuille d'assurés en rente d'éducation se renouvelle fréquemment, son taux technique moyen est proche des niveaux actuels des taux d'intérêts. Plus généralement, la distribution des taux techniques pour les deux garanties permet de tirer des conclusions quant aux générations des assurés :

Taux technique i	**% en rente de conjoint**	**% en rente d'éducation**
$0\% \leq i < 1\%$	9%	21%
$1\% \leq i < 2\%$	8%	30%
$2\% \leq i < 3\%$	22%	39%
$i \geq 3\%$	**61%**	**10%**

Table 2.11. Comparaison des distributions des taux techniques pour les rentes

Nous constatons d'une part que la moyenne du taux technique en rente de conjoint est de 2.9% contre 1.8% pour les rentes d'éducation. D'autre part, le pourcentage des rentiers ayant un taux technique élevé est plus important en rente de conjoint, en ce sens, plus de la moitié des conjoints bénéficiaires possèdent un taux technique supérieur à 3%.

Temporaire décès (garantie en capital)

Cette garantie offre le versement d'un capital en cas de décès de l'assuré survenant pendant la période de validité du contrat. Si l'assuré est en vie au terme de cette période, le contrat d'assurance prend fin. Les provisions sur ce périmètre (cf. 2.12) correspondent uniquement à des provisions pour sinistres à payer. Elles sont constituées pour payer les prestations relatives aux décès survenus avant la date d'évaluation.

	Provisions totales	PM	PSAP
Temporaire décès	144 895	0	144 895

Table 2.12. Provisions Solvabilité I - Temporaire décès

Santé

Cette garantie est obligatoire depuis l'entrée en vigueur de l'Accord National Interprofessionnel au 1er janvier 2016, elle permet de compléter les prestations santé de la sécurité sociale. De la même manière, seules des PSAP sont constituées sur cette garantie non-vie.

	Provisions totales	PM	PSAP
Santé	245 847	0	245 847

Table 2.13. Provisions Solvabilité I - Santé

Primes futures

Comme mentionné précédemment, la projection des contrats en Solvabilité II prévoit d'inclure les cotisations de l'année N+1. **Le montant total de primes** projetées pour le calcul du Best Estimate est de **1 120 737**, avec la répartition suivante par garantie :

Garantie	Cotisations
Arrêt de travail	249 874
Santé	645 853
Décès	225 010

Table 2.14. Primes N+1 projetées en lien avec la frontière des contrats Solvabilité II

2.5 Récapitulatif des variables d'intérêt par garantie

Le tableau suivant donne une synthèse des variables impactant chaque garantie compte-tenu de la modélisation et des hypothèses retenues pour notre étude. Certains paramètres classés sans impact peuvent avoir une influence sur les cashflows dans le cadre d'une approche de modélisation différente de celle retenue pour notre étude. À noter toutefois que la méthodologie d'agrégation qui sera proposée peut être généralisée à n'importe quel modèle de projection.

Garantie	Paramètres impactant la projection des flux futurs probables	Paramètres sans impact compte-tenu des hypothèses et de l'approche de modélisation
Incapacité / Invalidité	- Âge de l'assuré - Date d'entrée en IT/IP	- Le sexe de l'assuré (étant donné que les lois de maintien et de décès retenues en IT/IP ne tiennent pas compte du sexe) - L'échéance de la garantie, identique pour l'ensembles des assurés
Maintien de la garantie décès	- Âge de l'assuré - Date d'entrée dans l'état - Etat initial (incapacité ou invalidité)	
Rente d'éducation	- Âge et sexe du bénéficiaire - Catégorie socio-professionnelle du parent assuré - Taux technique - Fréquence de la rente (trimestrielle ou mensuelle)	- Terme de la rente, fixé à 28 ans
Rente de conjoint	- Âge et sexe du bénéficiaire - Taux technique - Fréquence de la rente	- Nature de la rente : viagère

Table 2.15. Synthèse des variables impactant la projection des cashflows

$\Longrightarrow$ Nous omettons les garanties temporaire décès et santé ainsi que les Model Points de primes futures car leur modélisation se fait par le biais des profils moyens. **Seules les garanties présentées dans le tableau ci-dessus entrent dans le périmètre de l'agrégation**. En revanche, l'impact sur le bilan économique sera mesuré au global de l'entité, en prenant en compte l'ensemble des garanties.

Chapitre 3

Modèle et hypothèses de projection

3.1 Présentation générale de l'outil

Comme mentionné précédemment, le calcul de la meilleure estimation des engagements de l'assureur *(Best Estimate Liabilities)* repose sur la projection des flux de trésorerie relatifs à l'ensemble des contrats en portefeuille.

Cette projection se fait grâce à un outil de modélisation actuarielle, les inputs du modèle correspondent aux données relatives aux assurés (Model Points de passif), les placements financiers ainsi qu'aux hypothèses techniques et économiques.

Avant toute chose, il convient de préciser que le pas de temps de la projection est mensuel et l'horizon de la projection est de 50 ans.

Par ailleurs, le modèle utilisé fonctionne par la méthode du Flexing, cela signifie que la modélisation se fait en deux étapes :

– **Une projection passif seul** permettant de déterminer les flux futurs liés uniquement au passif. Plus précisément, en sortie de cette projection, nous avons la chronique des prestations futures probables ainsi celle des provisions projetées ;

– **Une projection ALM** visant à modéliser l'interaction actif-passif. Elle prend en entrée les données des placements financiers de l'entité modélisée, ainsi que les sorties de la projection passif seul. Cette projection permet par ailleurs de modéliser les mécanismes de participation aux bénéfices.

D'autre part, pour chaque étape de modélisation (i.e. passif seul et ALM), le modèle permet de projeter les flux futurs aussi bien en situation centrale que dans les scénarios de chocs. Cela signifie que nous pouvons évaluer les variations du Best Estimate et de la valeur de marché des placements suite à l'application des sensibilités techniques et financières proposées par la formule standard de calcul du capital de solvabilité requis *(SCR)*. Un module de consolidation est donc prévu à ce titre, pour consolider les scénarios de choc

et pour appliquer les coefficients de corrélation qui existent entre les différents risques. En somme, le modèle de projection permet de construire le bilan économique et de calculer le ratio de solvabilité à la date évaluation.

En résumé, nous pouvons représenter schématiquement l'outil et la procédure de modélisation utilisés de la manière suivante :

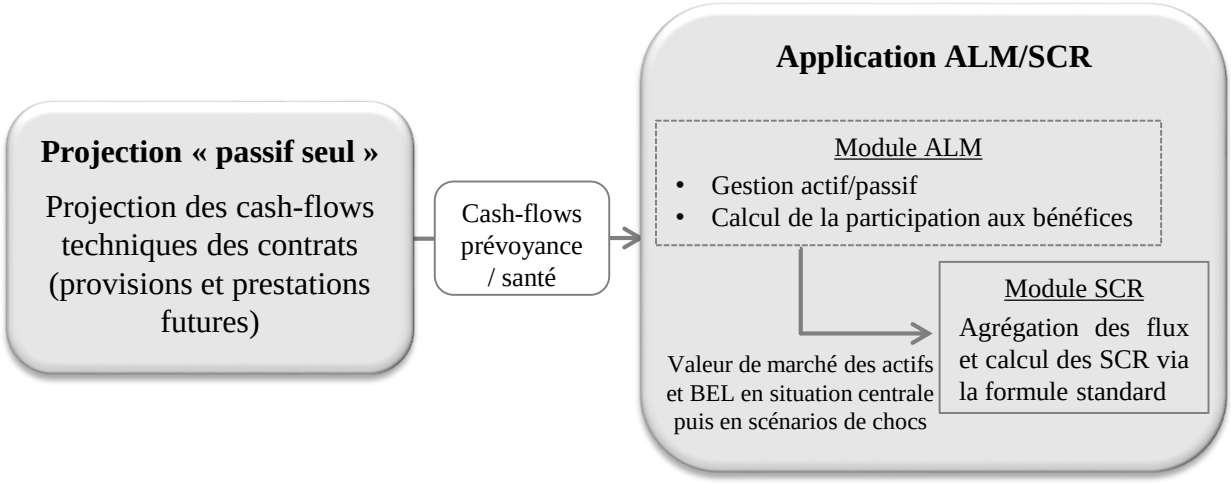


Schéma 3.1. Modèle de projection actuarielle

Lors de la mise en place de notre méthode, nous nous appuyons dans un premier temps sur les cashflows en sortie du modèle de la projection passif seul afin d'agréger le portefeuille de contrats. Ensuite, nous mesurons les impacts sur les indicateurs Solvabilité II et sur le bilan économique grâce à la projection ALM/SCR.

3.2 Approche de modélisation

La présente section a pour objectif de décrire l'approche de modélisation employée pour projeter les cashflows futurs liés à notre portefeuille de contrats. De ce fait, nous allons d'abord faire un point détaillé sur la modélisation du passif, avant d'évoquer la modélisation de l'actif, des mécanismes de participation aux bénéfices ainsi que de la stratégie de gestion actif-passif. Étant donné que notre étude a pour ambition de proposer une méthodologie d'agrégation du passif, seule la partie portant sur la modélisation du passif en prévoyance et santé sera présentée en détail. Les autres aspects de modélisation sont abordés brièvement dans le but de refléter les impacts de l'agrégation du passif sur le bilan économique.

Modélisation du passif

La modélisation des cashflows de trésorerie associés aux trois grands risques (Arrêt de travail, Santé et Décès) se fait selon deux approches de modélisation distinctes. L'objectif de cette partie est de présenter chacune des deux approches et de préciser leurs périmètres d'application en lien avec le portefeuille étudié.

Approche Markovienne

Cette approche de modélisation tient son nom des chaines de Markov en probabilité : il s'agit de processus dont le futur ne dépend du passé qu'à travers le présent. En d'autres termes, la probabilité pour qu'un assuré valide décède ne tient pas compte du passé de l'assuré : qu'il ait été incapable quelques mois auparavant ou non n'influence pas sa probabilité de décès en tant que valide.

Les quatre états représentés ci-dessous illustrent les évolutions possibles d'un assuré au cours du temps.

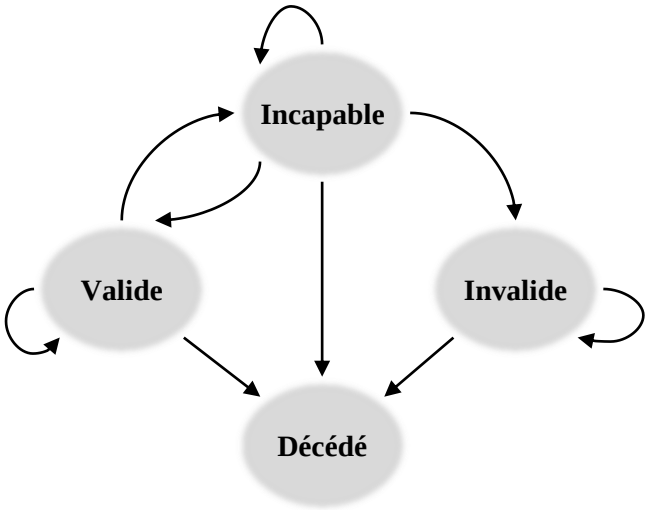


Figure 3.1. Transitions possibles entre les différents états de l'assuré

Le cœur de la modélisation Markovienne consiste à simuler, pour un état initial donné, l'ensemble des transitions probables d'une tête assurée vers l'un des états représentés ci-dessus. A cet effet, cette approche de modélisation permet :

- De projeter le nombre d'effectifs de chaque état
- De calculer les flux liés à une transition de la tête assurée vers un nouvel état ou de son maintien dans l'état initial (flux de prestations et de frais)
- D'évaluer les provisions de sinistres (PM et PSAP)

En guise d'exemple, la prestation en $t+1$ d'un assuré en invalidité en t correspond à sa probabilité de maintien en invalidité à la date $t+1$, multipliée par le forfait mensuel prévu par le contrat. La provision mathématique s'obtient alors en actualisant les prestations futures probables au taux technique en vigueur.
Pour les rentes de conjoint et les rentes d'éducation, le calcul s'applique plutôt au bénéficiaire désigné par l'assuré. Dans ce cas, cela consiste simplement à modéliser les probabilités de décès d'un rentier conjoint/enfant.
Finalement, les probabilités de transitions entre les différents états sont fournies par les tables réglementaires ou des tables d'expérience quand une loi a pu être calibrée sur les données du portefeuille. Étant donné que le pas de temps de la projection est mensuel, les taux calculés à partir des tables disponibles uniquement à la maille annuelle sont mensualisés.

L'approche Markovienne s'applique sur l'ensemble des Model Points renseignés en tête par tête. Dans le cadre de notre étude, cela concerne les assurés dont l'état initial est celui d'incapacité ou d'invalidité, ainsi que les bénéficiaires d'une rente d'éducation ou d'une rente de conjoint. Pour les Model Points relatifs au décès, à la santé et aux primes futures, une autre approche de modélisation est utilisée. Cela sera l'objet de la section suivante.

Approche S/P

Contrairement à l'approche décrite dans la section précédente, l'approche S/P consiste à suivre en moyenne l'évolution des grandeurs projetées (provisions, prestations, frais, commissions, etc.), sans qu'un décompte des effectifs ne soit effectué. Cette méthode permet ainsi, à partir d'une information limitée (chiffre d'affaires, ratio S/P et cadences de liquidation) de projeter les stocks et les flux futurs.

Les Model Points non sinistrés (i.e. relatifs aux primes futures) sont projetés à l'aide de l'approche S/P, en suivant les deux étapes suivantes :

- Première étape : à chaque encaissement de primes, calcul de la charge ultime de sinistralité associée. Cette charge de sinistralité est évaluée via l'application d'un ratio S/P au chiffre d'affaires projeté
- Deuxième étape : la charge de sinistralité calculée est écoulée dans le temps selon des cadences de règlement définies en hypothèses

Ainsi, le nombre de Model Points de primes futures est conditionné par la maille de calcul des S/P et des cadences de règlement.

Par ailleurs, les provisions pour sinistres à payer (PSAP) relatives aux Model Points sinistrés sont également écoulées selon une hypothèse de cadences de règlement applicable au stock de contrats. En effet, les prestations sous-jacentes peuvent avoir lieu sur plusieurs mois/années en fonction de la garantie souscrite.

En définitive, le tableau ci-dessous fournit une synthèse de l'approche de modélisation retenue selon la nature du Model Point et en fonction du type de la provision constituée.

Model Point	Garanties	PM	PSAP	Approche de modélisation
Sinistré *(données tête par tête)*	Arrêt de travail, rente d'éducation, rente de conjoint	✔	✔	Approche Markovienne pour les PM et cadences d'écoulement du stock pour les PSAP
Sinistré *(profils moyens)*	Santé, temporaire décès	✘	✔	Cadences d'écoulement du stock
Non sinistré *(profils moyens)*	Toutes garanties	S/P pour estimer la charge ultime de sinistralité puis calcul des prestations futures via les cadences d'écoulement relatives aux primes futures		

Table 3.1. Synthèse des approches de modélisation par catégorie de Model Point

Autres aspects de modélisation

Modélisation de l'actif

Cette partie présente brièvement la projection des catégories d'actifs *(Equity et Bond)* composant le bilan de l'entité étudiée. À noter que la projection de l'actif repose avant tout sur le scénario économique à la date d'évaluation et particulièrement sur courbe de taux fournie par l'EIOPA.

En premier lieu, la catégorie *Equity* comprend les actifs issus des classes actions, immobilier et OPCVM (environ 25% du portefeuille étudié).
L'évolution de la valeur de marché des titres sous-jacents est modélisée au moyen d'un indice de performance nette, il est défini comme étant la différence entre :

- Le taux de performance brut du marché paramétré dans le scénario économique de projection ;
- Le taux de distribution de richesse (dividende dans le cas d'une action ou loyer dans le cas d'un bien immobilier).

En second lieu, la catégorie *Bond* comprend les titres obligataires. D'un point de vue modélisation, leur valeur de marché est calculée à chaque pas de temps en actualisant les flux futurs (coupons et nominal à l'échéance).
Finalement, il convient de noter que les produits financiers projetés sont composés d'une part des coupons générés par les titres obligataires et d'autre part des dividendes/loyers

générés par l'*Equity*.

Stratégie de gestion actif-passif

L'évolution de la trésorerie est au cœur des interactions Actif-Passif. En effet, la projection de l'actif donne lieu à l'encaissement/décaissement de flux tels que les revenus issus des dividendes, tandis que la projection du passif induit la survenance de flux tels que l'acquisition de primes ou le paiement des prestations. Ces flux sont comptabilisés au débit ou au crédit de la trésorerie.

Une stratégie d'investissement est mise en place dans le modèle de projection de manière à aligner le portefeuille d'actif à l'allocation cible définie au niveau de l'entité. Ce rebalancement a également un impact sur la trésorerie.

Le niveau de cash après rebalancement des actifs doit permettre de financer les garanties réglementaires et contractuelles. Si tel n'est pas le cas, alors une stratégie de revalorisation est mise en œuvre. Cette stratégie définit le montant de plus-values à réaliser afin d'obtenir le taux optimal de rendement des actifs.

Mécanismes de participation aux bénéfices

Quand une clause de participation aux bénéfices est prévue dans le contrat collectif, un compte de participation aux bénéfices est établi entre l'entité et le client. Ce compte a pour finalité la constitution puis la dotation/reprise de la provision d'égalisation et de la réserve générale. L'ensemble des clauses régissant le compte de participation aux bénéfices est défini dans un protocole client.

A la suite de la projection actif-passif, le solde technique du compte client est évalué. Le mécanisme de participation aux bénéfices fonctionne globalement de la manière suivante :
– Si le solde est positif, alors la provision d'égalisation est dotée à hauteur de 75% du résultat, l'excédent est alors mis en réserve générale (provision non déductible).
– Si le solde est négatif alors la réserve générale est reprise en priorité.

Par ailleurs, le stock de provision d'égalisation est plafonné à hauteur d'un pourcentage déterminé en fonction des effectifs assurés au sein du contrat collectif. Ce pourcentage est fixé à 60% pour les contrats du portefeuille étudié.

Finalement, il convient de noter que l'ensemble des provisions sont dotées d'une rémunération financière sur la base d'un indice convenu contractuellement. Dans le cadre de ce mémoire, nous faisons l'hypothèse que la rémunération des provisions se fait à 90% du taux de rendement actuariel (TRA) projeté.

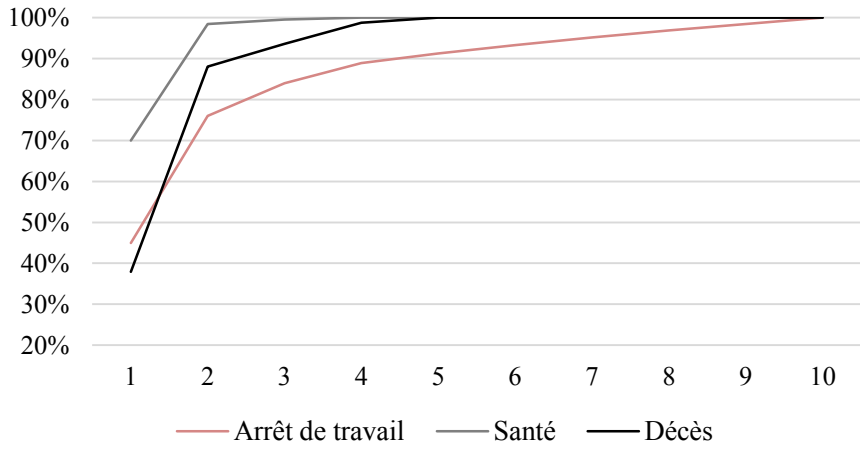
3.3 Hypothèses retenues

L'objectif de cette section est de présenter les hypothèses permettant d'évaluer les prestations futures probables et de projeter les provisions en stock (PM et PSAP).

Il convient dans un premier temps de préciser les lois utilisées dans le cadre de la projection des prestations futures probables selon l'approche Markovienne :

- Pour les probabilités de maintien en incapacité/invalidité, nous utilisons les **tables du BCAC publiées en 2010**. Cette loi est indépendante du sexe de l'assuré. Il en est de même pour le passage de l'incapacité temporaire à l'invalidité permanente ;
- Pour le décès des assurés dont l'état initial est celui d'incapacité ou d'invalidité, nous avons recours aux tables **du BCAC publiées en 2002** relatives à la mortalité des personnes en arrêt de travail, toujours sans distinction en fonction du sexe ;
- Les probabilités de décès en rente d'éducation et en rente de conjoint sont calculées sur la base des tables générationnelles réglementaires, **TGH-05** pour les hommes et **TGF-05** pour les femmes ;
- Finalement, deux **lois d'expérience pour la poursuite des études** en rente d'éducation sont calibrées, en fonction de la catégorie socio-professionnelle du parent (cadre/non cadre).

Par ailleurs, l'hypothèse d'écoulement des provisions pour sinistres à payer relatives au stock de contrats est calibrée par risque, comme le montre le graphe ci-après.



Graphique 3.1. Cadences d'écoulement des provisions pour sinistres à payer en stock

D'autres hypothèses seront requises pour évaluer le Best Estimate, notamment des hypothèses de frais, de sinistralité future et de participation aux bénéfices. Ces dernières sont présentées dans la section 6.1 portant sur le calcul du BEL. Finalement, les hypothèses financières retenues s'appuient sur les données communiquées par l'EIOPA au 31/12/2018.

Chapitre 4

Méthodes de Data Mining

Dans ce chapitre, nous abordons deux grandes familles de méthodes Data Mining qui nous seront utiles lors de la mise en place de notre méthode d'agrégation du passif. La première est la classification non-supervisée : cela nous permettra de créer des individus représentatifs en s'appuyant sur les flux associés aux assurés dans notre base d'apprentissage. La deuxième famille de méthodes de classification supervisée sera utilisée pour établir les liens entre les caractéristiques des assurés et leurs groupes représentatifs. Les notations utilisées dans ce chapitres sont résumées dans le schéma suivant :

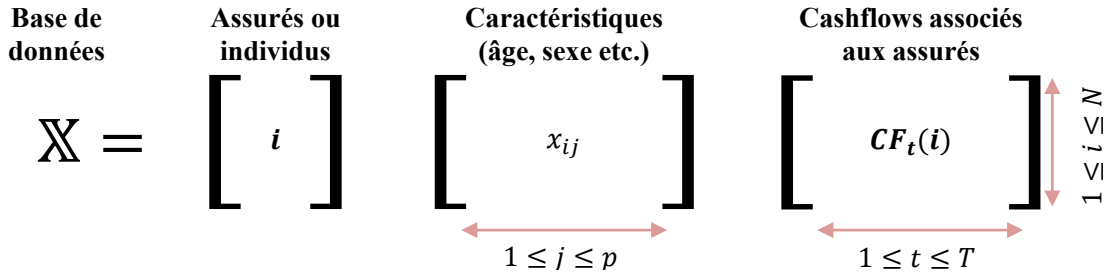


Schéma 4.1. Base de données des assurés et notations retenues

4.1 La classification non-supervisée

Les méthodes de classification non-supervisée permettent, à partir d'une base de données d'individus caractérisés par des variables quantitatives, de former des groupes homogènes d'individus et ce dans l'optique d'agréger et de structurer la base de données.
Soit $\mathbb{X}$ une base de données contenant N individus. L'idée de la classification non-supervisée consiste à considérer que les N individus de la base de données manipulée sont issus de $K \leq N$ populations.

Les règles d'appartenance d'un individu $i \in [1, N]$ à une population $k \in [1, K]$ n'étant pas connues a priori par l'algorithme, elles sont définies sur la base de la similarité qui peut exister entre les valeurs prises par variables quantitatives associées à chaque individu. Dans le cadre de notre étude, les variables quantitatives considérées correspondent aux cashflows associés aux assurés.

4.1.1 Panorama et comparaison des méthodes

Il existe plusieurs méthodes de classification non-supervisée se distinguant principalement par :

- L'algorithme **itératif** ayant permis d'obtenir les classes homogènes ;
- La **structure** de la classification obtenue (hiérarchie, partition, etc.) ;
- Approche sous-jacente **(géométrique ou probabiliste)**.

Nous nous intéressons à quatre méthodes classiques (K-means, Classification Ascendante Hiérarchique, DBSCAN et Modèle de Mélange Gaussien). D'abord, nous présentons très brièvement le principe de chaque méthode, ensuite nous comparons leur efficacité compte-tenu des données et des ressources informatiques à disposition. Finalement, la formulation mathématique de la méthode retenue sera présentée, tout en étudiant les paramètres entrant en jeu dans le processus de classification.

DBSCAN

La méthode DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est due à [Ester 1996]. Son principe consiste à créer, au sein de la base de données initiale, des sous-groupes à forte densité. Cette méthode fait appel à deux paramètres. Un premier paramètre de distance ϵ, permettant de définir le rayon de recherche des individus les plus proches. Le deuxième étant le nombre minimal d'individus d'une classe p_{min}. Autrement dit, le cardinal des groupes créés doit être au moins égal à ce nombre minimum de points. Cela conduit donc à trois catégories d'individus représentées dans la figure 4.1 :

- *Core Points* : les points dont le voisinage de rayon ϵ inclut au moins p_{min} individus ;
- *Border Points* : les points dont le voisinage de rayon ϵ contient un nombre d'individu inférieur à p_{min} ;
- *Noise Points* : les points dont le voisinage de rayon ϵ est vide.

L'algorithme de classification parcourt l'ensemble des individus de la base, en intégrant au sein du même groupe les individus appartenant au $\epsilon - voisinage$ de chaque point, sous contrainte de former des groupes constitués au minimum de p_{min} éléments.
En conséquence, la méthode DBSCAN ne met pas nécessairement l'ensemble des individus dans une classe homogène : c'est le cas des Noise Points qui n'appartiennent à aucun groupe.

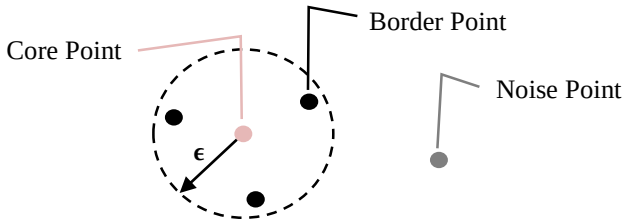


Figure 4.1. Illustration de l'algorithme de classification DBSCAN (*avec* $p_{min} = 4$)

K-means

La deuxième méthode que nous abordons est celle des K-means (ou K-moyennes). Comme son nom l'indique, cette méthode a pour but de créer une partition à K classes de sorte que les individus au sein de chaque groupe soient centrés autour de sa moyenne géométrique i.e. de son barycentre.

La méthode repose sur un algorithme itératif et sur une mesure de distance entre deux individus. Elle consiste à réitérer les étapes suivantes afin de minimiser un critère mesurant la qualité du partitionnement : le critère utilisé est l'inertie intra-classe (cf. section 4.1.2).

- Initialisation : K éléments sont tirés aléatoirement dans la base de données à segmenter, ces derniers correspondent dans un premier temps aux centres des classes ;
- Affectation : l'algorithme parcourt l'ensemble des individus et les affecte à la classe dont le barycentre est le plus proche ;
- Représentation : le barycentre des classes est recalculé à chaque allocation d'un individu à une classe.

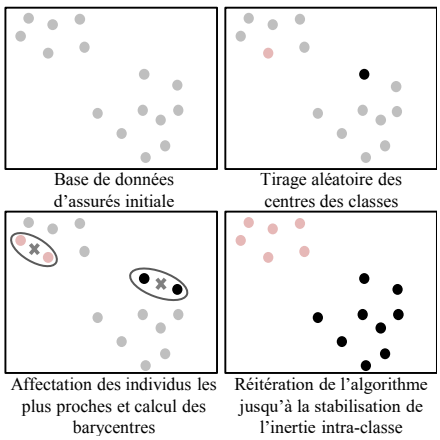


Figure 4.2. Illustration de la méthode des K-means (partitionnement en deux classes)

Classification ascendante hiérarchique

Plutôt que de fournir une seule partition de K éléments, la classification ascendante hiérarchique (CAH) réalise des partitions emboîtées constitués de $N, N-1, N-2, ..., 2$ éléments. Le résultat de l'algorithme est représenté graphiquement par un dendrogramme (cf. figure 4.3). Ceci dit, construire une partition de K classes revient à réaliser une coupure au niveau adéquat i.e. le niveau conduisant à K intersections entre l'axe de coupure et les branches séparant chaque classe.

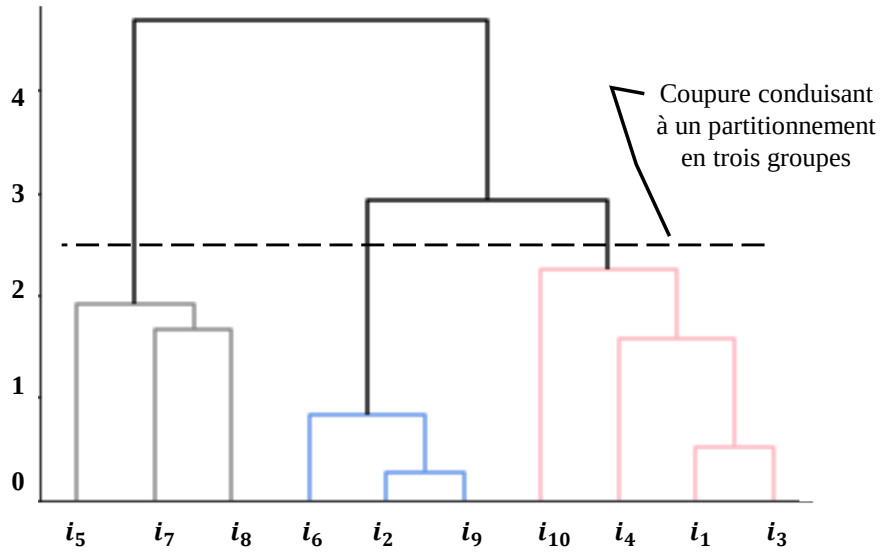


Figure 4.3. Illustration du dendrogramme obtenu en appliquant la classification ascendante hiérarchique sur 10 individus indexés par $i_{1 \leq l \leq 10}$

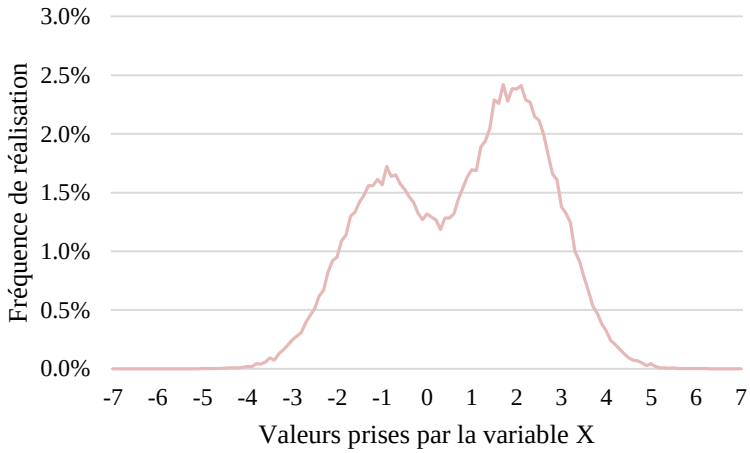
Pour obtenir cette classification, l'algorithme construit dans un premier temps un partitionnement à N singletons, à partir des N individus disponibles dans la base des données à partitionner.
Ensuite, l'algorithme calcule la matrice des distances entre les individus.
Finalement, les partitions à $N-1, N-2, ..., 2$ classes sont créées de proche en proche en incluant dans chaque classe l'individu ayant la plus petite distance à la classe en question.

Ainsi, cette méthode requiert de définir non seulement une mesure de distance entre deux individus, mais aussi une mesure de distance entre une classe et un individu.

Modèle de Mélange Gaussien

Pour cette méthode appelée communément GMM (Gaussian Mixture Model), la similarité entre deux individus ne s'appuie pas sur des mesures géométriques mais plutôt sur des considérations probabilistes. En d'autres termes, deux individus appartiennent à la même classe s'ils sont issus de la même distribution de probabilité.

Prenons l'exemple simple d'une base de données d'individus caractérisés par une variable X dont l'histogramme des valeurs est donné par le graphique 4.1.



Graphique 4.1. Illustration du Modèle de Mélange Gaussien

La forme de la densité suggère que la variable X est issue d'une combinaison des deux variables aléatoires Gaussiennes $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$. De ce fait, l'objectif de cette méthode de classification est de calculer la probabilité que chaque observation corresponde à la réalisation d'une des deux variables aléatoires Gaussiennes, ce qui conduit à segmenter notre base de données en deux groupes homogènes.

Pour ce faire, la première étape consiste à initialiser les paramètres des lois recherchées μ et σ ainsi que la proportion λ de la population issue de chaque groupe, par exemple en s'appuyant sur la forme de la densité observée. Ensuite, l'algorithme de classification consiste à itérer les deux étapes suivantes :

(1) Espérance : calcul des probabilités d'appartenance à chaque groupe, compte-tenu de la valeur des paramètres μ, σ et λ ;

(2) Maximisation : réévaluation des paramètres suite au calcul des probabilités à l'étape (1) en maximisant la vraisemblance.

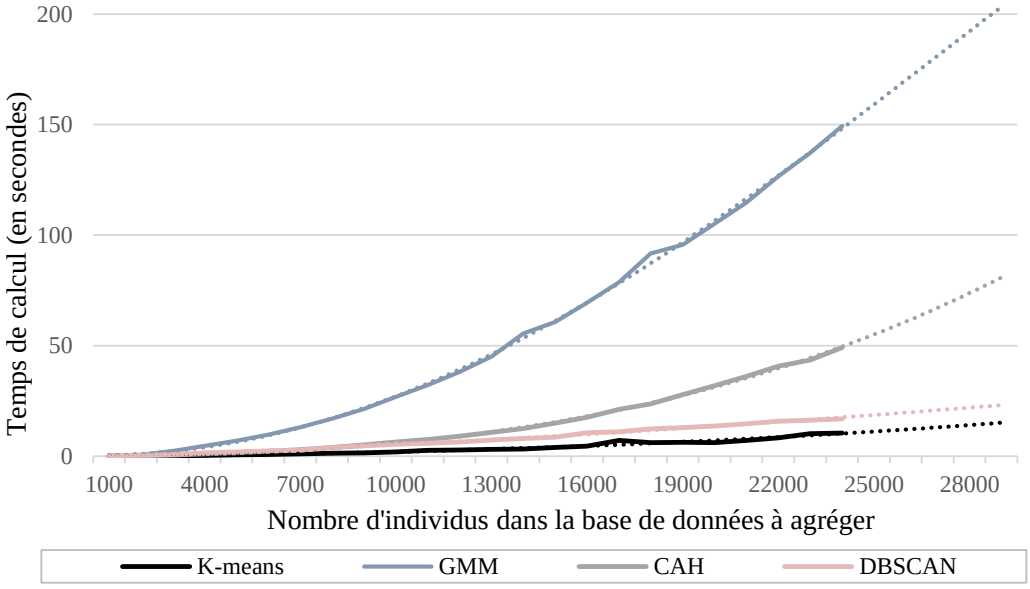
Finalement, une condition d'arrêt de l'algorithme itératif intervient quand les paramètres recalculés deviennent stables.

Comparaison des méthodes et choix de la méthode adaptée à la problématique

Avant de choisir la méthode d'agrégation à utiliser, il convient de donner quelques précisions concernant les dimensions des bases de données qui seront manipulées dans le cadre de notre étude.

D'une part, des bases d'apprentissage seront construites pour chaque garantie offerte aux assurés : elles contiennent entre 10 000 et 70 000 individus. D'autre part, l'agrégation du passif se fera sur la base des flux futurs probables relatifs à chaque assuré présent dans ces bases d'apprentissage. Ainsi, le nombre de variables caractérisant les individus est lié à l'horizon de projection retenu.

Dès lors, la complexité temporelle des méthodes (cf. annexe) est cruciale surtout qu'il sera question de réaliser plusieurs de tests avant d'obtenir l'agrégation optimale du passif. Un comparatif du temps de calcul des différentes méthodes a été réalisé, en prenant en compte les propriétés des bases de données manipulées. Nous construisons en premier lieu des bases de données synthétiques caractérisées par 50 variables quantitatives (horizon de la projection) et dont le nombre d'individus varie entre 1 000 et 25 000. Ensuite, à l'aide de la libraire *Scikit-learn* du langage Python, nous effectuons des agrégations conduisant à réduire le nombre d'individus à hauteur de 10%. Enfin, nous mesurons le temps de calcul pour chaque méthode ainsi que pour les différentes bases de données considérées. Les résultats sont présentés dans le graphe ci-dessous :



Graphique 4.2. Comparaison du temps de calcul des méthodes de classification non-supervisée *(les courbes en pointillés représentent les tendances polynomiales du temps de calcul)*

Sur la base de la comparaison réalisée, nous pouvons conclure que le temps de calcul est relativement faible et qu'il est quasiment identique pour les différentes méthodes, à l'ex-

ception de la méthode GMM pour laquelle le temps de calcul augmente considérablement avec la taille de la base à agréger. Par ailleurs, il convient de mentionner que :

- La classification ascendante hiérarchique (CAH) nécessite de calculer la matrice des distances entre les individus au sein des bases de données. Cela ne pose pas de problème en terme de temps de calcul, toutefois, l'application de cette méthode sur les bases de données manipulées conduit à la saturation de l'espace mémoire des ressources informatiques à disposition ;
- L'avantage principal de la méthode DBSCAN est sa robustesse face aux valeurs aberrantes (*outliers*). Dans le cadre de notre étude, les bases de données sont générées de sorte qu'il y ait très peu de valeurs aberrantes, cet aspect est d'autant plus important qu'il permettra d'assurer l'efficacité des méthodes d'apprentissage automatisé utilisées dans un deuxième temps lors de la mise en place de notre méthode d'agrégation. À noter que la méthode DBSCAN nécessite de fixer au préalable deux paramètres ϵ et p_{min} alors que les K-means prend en input un paramètre simple qui est le nombre de groupes à construire.

Enfin, la méthode des **K-means** est largement documentée et sa stabilité a été éprouvée dans plusieurs domaines de recherche, nous optons pour cette méthode pour la suite de notre étude.

4.1.2 Formulation mathématique de la méthode retenue

L'objectif de cette section est de poser la formulation mathématique de la méthode des K-means introduite par [MacQueen 1967] ainsi que de présenter succinctement les notions sur lesquelles elle repose.

Soit $\mathbb{X}$ une base de données de N individus, caractérisés par T variables quantitatives, qui correspondent dans notre cas aux cashflows associés aux assurés :

$$\mathbb{X} = CF_t(i), 1 \leq i \leq N \text{ et } 1 \leq t \leq T$$

L'objectif de l'algorithme est de construire une partition $P = (C_1, C_2, ..., C_K)$ à K classes, d'intersection vide et dont l'union corresponde à la base totale $\mathbb{X}$. Pour construire cette partition, nous utilisons l'algorithme itératif décrit précédemment. Cependant, nous avons besoin de définir deux éléments :

1– Une mesure de distance entre les individus : les assurés sont caractérisés par leur flux futurs probables. Dans le cadre de notre étude, nous aurons recours à la distance Euclidienne (ce choix sera détaillé dans le chapitre 7). De ce fait, pour deux assurés i_1 et i_2 la distance est donnée par :

$$D(i_1, i_2) = ||CF(i_1) - CF(i_2)||_2 = \sqrt{\sum_{t=1}^{T} (CF_t(i_1) - CF_t(i_2))^2} \tag{4.1}$$

Où T représente l'horizon de projection des cashflows.

2– Un critère d'homogénéité des classes i.e. une mesure de la qualité du partitionnement. Le critère couramment utilisé est celui de l'inertie intra-classe W. Pour la partition P, l'inertie intra-classe est donnée par :

$$W_P = \sum_{k=1}^{K} I(C_k) \tag{4.2}$$

Avec $I(C_k) = \sum_{i \in C_k} D^2(i, g_k)$ l'inertie de la classe C_k de barycentre g_k.

Par ailleurs, l'inertie totale W de la base d'individus est donnée selon la même formule. En posant g le barycentre de $\mathbb{X}$, nous avons :

$$W = \sum_{i=1}^{N} D^2(i, g) \tag{4.3}$$

Le ratio $R_P = 1 - \frac{W_P}{W}$ correspond à l'inertie expliquée d'une partition : il varie entre 0% et 100% et augmente avec le nombre de classes de la partition. Il atteint sa valeur maximale quand la partition P est constituée de N singletons, sa valeur minimale est atteinte pour la partition triviale constituée d'un seul élément.

De ce fait, pour deux partitions à K classes P_1 et P_2. On considère que la partition P_1 est de meilleure qualité que P_2 au sens de l'inertie expliquée si $R_{P_1} \geq R_{P_2}$. L'inertie totale étant indépendante du partitionnement choisi, cela revient à dire que $W_{P_1} \leq W_{P_2}$

La détermination de la meilleure partition à K classes se résume alors à un problème d'optimisation : il s'agit de minimiser la grandeur W_{P_K} sur l'ensemble des partitions P_K possibles.

L'algorithme des K-means décrit précédemment conduit à un minimum local de l'inertie intra-classe, cela signifie que le partitionnement obtenu après convergence de l'algorithme dépend du choix des centres initiaux des groupes. Une stratégie classique pour remédier à cette limite consiste à lancer l'algorithme plusieurs fois avec des configurations différentes et de prendre la partition ayant la plus petite inertie intra-classe. De surcroît, il existe des techniques permettant d'optimiser l'initialisation de l'algorithme, ces dernières seront présentées dans la section suivante.

4.1.3 Choix des paramètres

Initialisation de l'algorithme

Pour chaque base de données manipulée, nous augmentons le nombre de centres initiaux jusqu'à la stabilisation du partitionnement obtenu (i.e. jusqu'à ce que l'augmentation du

nombre de centres initiaux induise très peu de changements dans la classification). De surcroît, à chaque étape, le choix des centres initiaux est optimisé par la méthode des K-means++ proposée par [Arthur 2007].

Au lieu de choisir K centres de groupes initiaux selon un tirage aléatoire uniforme, l'idée consiste à tirer aléatoirement un seul centre, puis réitérer les deux étapes suivantes jusqu'à l'obtention de K centres :

(1) Calcul des distances $d_x = D(i_0, x)$ entre le dernier centre choisi i_0 et l'ensemble des individus x de la base ;

(2) Tirage d'un nouveau centre selon une loi de probabilité qui croit en d_x^2 : l'objectif étant d'initialiser l'algorithme avec des centres éloignés.

En conséquence, cette méthode d'initialisation permet non seulement d'optimiser le choix des centres mais aussi de réduire le nombre d'itération nécessaires pour la convergence de l'algorithme des K-means.

Nombre optimal de classes

Comme évoqué précédemment, notre méthode de partitionnement prend en paramètre le nombre de classes K à former à partir des N individus à disposition. Il existe plusieurs critères permettant de choisir le nombre de classes optimal. Nous pouvons citer notamment la méthode d'interprétation graphique des Silhouettes, s'appuyant sur la comparaison entre les distances des individus à leurs classes d'appartenance (cohésion) et leur distance aux autres classes (séparation).

Dans le cadre de notre étude, le choix du nombre de classes se fera sur la base d'un critère adapté à la nature de la problématique étudiée. Plus précisément, il s'agira d'évaluer l'impact de l'agrégation sur les cashflows futurs. Ce critère de validation sera présenté en détail dans le chapitre 5.

Individu représentatif d'une classe

Pour finir cette section, il convient de définir une notion qui nous sera utile lors de l'application de notre méthode d'agrégation du passif.

A la suite de l'application des K-means sur notre base d'assurés, nous aurons à disposition un partitionnement de la base en K classes. Chaque classe possède un barycentre calculé en fonction des cashflows induits par les individus au sein du groupe.

Dans certains cas, la chronique des cashflows du barycentre ne correspond pas forcément à un assuré appartenant au groupe. Par exemple, pour un groupe contenant 2 assurés ayant les vecteurs de flux CF et CF', le barycentre du groupe est un assuré synthétique, n'appartenant pas à la base initiale et ayant le vecteur de flux $\frac{CF+CF'}{2}$. De plus, la définition des caractéristiques (âge, sexe, ancienneté, etc.) conduisant à une chronique de cashflows

identique à celle du barycentre s'avère très complexe. En effet, cela reviendrait à trouver la réciproque de la fonction qui fait correspondre les caractéristiques des assurés à leur cashflows futurs.

Pour les raisons citées précédemment, une classe sera représentée par l'individu dont la chronique des cashflows est la plus proche au barycentre. Autrement dit, pour une classe C_k de barycentre g_k, l'individu représentatif i_k de la classe est obtenu par la formule suivante :

$$i_k = \underset{i \in C_k}{\operatorname{argmin}} D(i, g_k) \tag{4.4}$$

4.2 La classification supervisée

On considère une base de données contenant N individus caractérisées par p variables quantitatives ou qualitatives, par exemple l'âge et le sexe des assurés. On suppose par ailleurs que chaque individu appartient à une classe $C_k \in \{C_1, C_2, ..., C_K\}$, connue a priori.

À noter que les classes sont déterminées grâce à la méthode présentée précédemment, en utilisant les cashflows associés aux assurés. L'objectif de la classification supervisée est d'établir les liens qui existent entre les caractéristiques des assurés et leur classe d'appartenance. Autrement dit, il s'agit de méthodes d'apprentissage automatisé, permettant de définir les règles de correspondance entre les assurés et leur classe afin de les généraliser à de nouveaux individus qui ne figurent pas dans notre base de donnée d'apprentissage.

De ce fait, ces méthodes seront utilisées dans le cadre de notre étude dans le but de généraliser, à un nouveau portefeuille, l'agrégation effectuée dans un premier temps par les méthodes de classification non-supervisée. Il s'agira donc de trouver, sur la base des caractéristiques telles que l'âge le sexe ou l'ancienneté, le groupe représentatif d'un assuré. L'utilisation de chaque famille de méthodes est synthétisée dans le schéma ci-après.

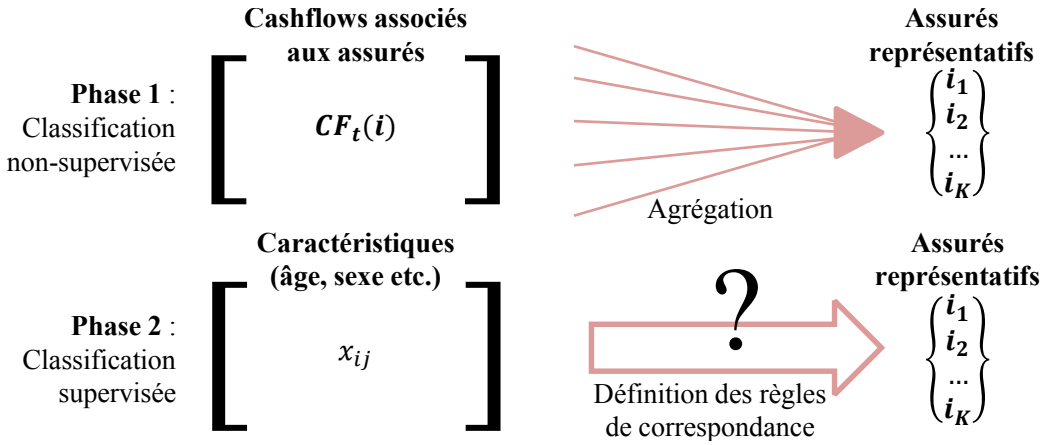


Schéma 4.2. Exploitation des méthodes de classification supervisée et non-supervisée

4.2.1 La problématique d'asymétrie des classes

La problématique principale à laquelle se heurtent les méthodes de classification supervisée est liée aux propriétés de la base de données d'apprentissage. En effet, un algorithme ou une méthode ne peut identifier les règles d'appartenance à une classe donnée que si l'on dispose d'un nombre d'individus suffisant pour cette classe. Ainsi, ce n'est pas uniquement la dimension totale de la base d'apprentissage qui nous intéresse, mais aussi le nombre d'assurés disponibles pour chaque classe.

Prenons l'exemple d'une classification de N individus portant sur une réponse binaire i.e. nous avons deux classes à prédire $K = 2$. Supposons en outre que 99% des individus disponibles concernent la classe C_0 et seules 1% concernent la classe C_1. Dans ce cas extrême, la fonction $i \mapsto C_0$ faisant correspondre n'importe quel individu i à la classe C_0 est bien ajustée sur les données de la base d'apprentissage i.e. elle prédit correctement 99% des individus, cependant elle n'est pas forcément adaptée à la distribution réelle des individus.

Pour répondre à cette problématique, nous nous inspirons de la technique de suréchantillonnage appelée SMOTE (cf. [Chawla 2002]). L'idée principale consiste à appliquer des perturbations sur la base de données initiale afin de générer de nouveaux individus permettant ainsi d'équilibrer les classes. Les modalités d'application des perturbations sur la base de données des assurés seront précisées dans le chapitre 5, en tenant compte de la nature des variables caractérisant les assurés.

4.2.2 Validation croisée et qualité de la prédiction

Avant de détailler les méthodes d'apprentissage automatisé qui seront utilisées dans le cadre de ce mémoire, il convient de préciser le critère retenu pour la validation de la qualité des prédictions réalisées.

Pour chaque garantie étudiée, nous disposons d'une base de données contenant les caractéristiques des assurés ainsi que le groupe auquel ils appartiennent. Une méthode communément utilisée consiste à diviser (selon un tirage aléatoire) cette base de données en deux parties : une partie pour l'apprentissage (i.e. la base permettant d'entraîner l'algorithme de classification supervisée) et une deuxième partie permettant de tester la précision de l'algorithme construit. En général, la base de test contient 25% des individus. Ensuite, la qualité de la prédiction est mesurée simplement en évaluant le rapport des individus correctement classées sur le nombre total d'assurés au sein de la base de test : cette mesure de la qualité de prédiction sera appelée **pourcentage d'adéquation**.

En revanche, de manière générale, diviser la base de données en deux parties présente quelques limites car cela implique que la validation soit réalisée avec un seul jeu de données. En d'autres termes, le score de prédiction ne sera pas suffisamment précis dès que la division

réalisée induit un changement dans la proportion de chaque classe d'individus.

En conséquence, nous aurons recours à la méthode de validation croisée des L-folds. Son application est assez simple : au lieu de diviser la base de données initiale en deux parties, l'idée consiste à réaliser cette démarche L fois. Autrement dit, nous fixons par exemple le paramètre L à 4 : la base de données est donc divisée par tirage aléatoire en 4 parties. Ensuite, les phases d'apprentissage et de test peuvent être itérées 4 fois en utilisant à chaque étape une des parties pour le test et son complémentaire pour l'apprentissage. Finalement, le score de prédiction correspond à la moyenne des pourcentages d'adéquation obtenus à chaque itération.

Finalement, il convient de mentionner que d'autres mesures de performance s'adaptant parfaitement à la problématique de la classification à classes multiples ont été proposées dans la littérature. Nous pouvons citer notamment l'indicateur proposé par [Cohen 1960], ce dernier est très utile lorsque la base de données manipulée présente un déséquilibre entre les classes (par exemple des classes sous-représentées au sein de la base de données). Dans le cadre de nos travaux, les scores de prédiction qui seront présentés correspondent à des pourcentages d'adéquation, il convient cependant de mentionner que les deux indicateurs ont été testés : en effet, ils fournissent des résultats très proches car les bases de données manipulées présentent un équilibre entre les classes, ce qui permet de rendre le pourcentage d'adéquation aussi précis que l'indicateur de [Cohen 1960] cité précédemment.

4.2.3 Les arbres de décision

Nous nous intéressons dans un premier temps aux arbres de décisions et particulièrement à la méthode Classification And Regression Trees introduite par [Breiman 1984].

Pour commencer, notons que cette méthode peut être utilisée aussi bien en classification qu'en régression (i.e. pour les problèmes de prédiction d'une variable quantitative). De manière générale, un arbre est composé de nœuds de décision (construits à partir des variables caractérisant les individus), et de feuilles contenant les classes C_k. La racine de l'arbre contient toujours la variable la plus discriminante, c'est-à-dire ayant le plus d'impact dans la détermination de la classe d'appartenance d'un individu.

Ensuite, dans le but de comprendre le mode de construction des arbres de décision, nous considérons un exemple simple qui s'approche davantage de notre problématique. Supposons que l'on dispose d'une base contenant 26 assurés caractérisés par leur âge et par l'ancienneté d'occurrence de leur sinistre. De plus, nous supposons que les 26 assurés appartiennent à trois classes, représentées par trois couleurs différentes dans la figure 4.4. Il convient d'attirer l'attention sur le fait que les classes d'appartenance des assurés ne sont pas obtenues sur la base de la proximité géométrique des individus. En effet, les classes sont obtenues en appliquant les méthodes de classification non-supervisée aux cashflows associés aux assurés.

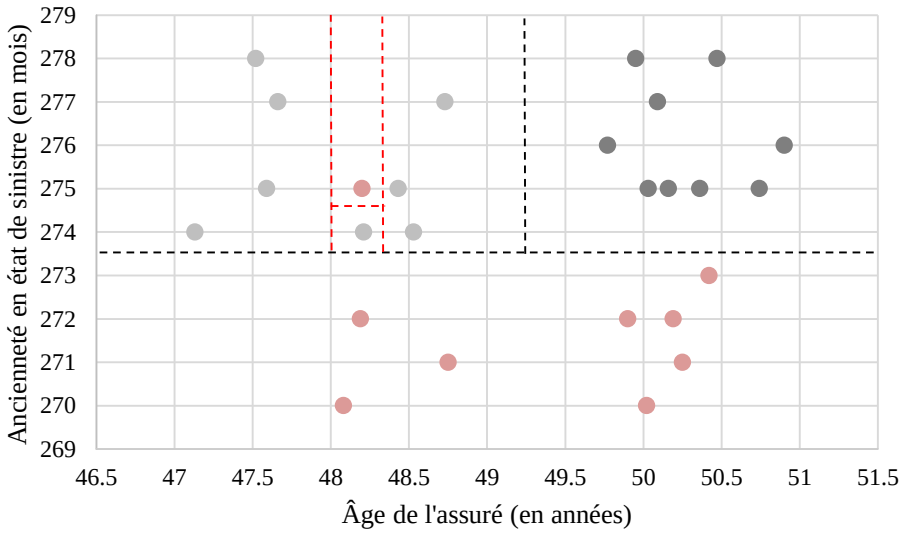


Figure 4.4. Extrait d'une base d'assurés utilisée en apprentissage automatisé

Pour traduire cette représentation en un arbre de décision, l'idée de base consiste à segmenter les variables caractérisant les assurés (cf. les lignes en pointillés) afin d'identifier les zones de l'espace contenant une classe unique. En s'appuyant sur le premier découpage (lignes en noir), nous obtenons l'arbre de décision présenté dans la figure 4.5.

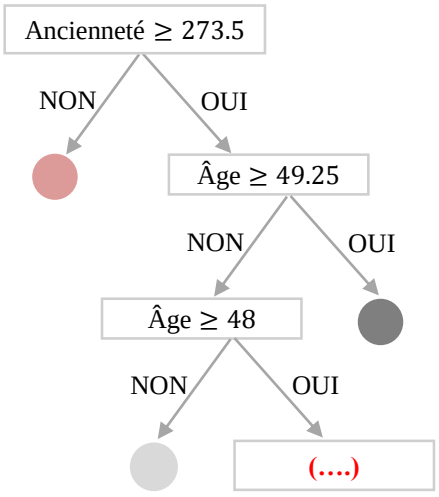


Figure 4.5. Illustration de l'arbre obtenu selon la première segmentation

Comme nous pouvons le constater, nous avons obtenu un arbre de décision qui s'adapte à toutes les observations de notre base de données, à l'exception d'une seule. Deux questions se posent donc à ce niveau :

1– Quel critère nous permet d'automatiser et d'optimiser ce processus de découpage des zones de l'espace ?

2– Faut-il s'arrêter à une profondeur donnée ou développer davantage les branches de l'arbre pour s'adapter complètement aux données de la base d'apprentissage ?

Gain d'information et entropie

La réponse à la première question est fournie par la théorie de l'information et plus précisément par l'entropie de Shannon. En effet, le découpage optimal des variables caractérisant les individus est celui qui maximise la fonction de gain d'information.

L'entropie est une mesure d'hétérogénéité. En d'autres termes, elle atteint sa valeur minimale quand les individus de la base de données appartiennent tous à la même classe. En revanche, sa valeur maximale est atteinte quand la distribution des classes est uniforme.

Pour chaque variable $x_i, 1 \leq i \leq p$ caractérisant les individus, l'arbre de décision est construit de manière récursive en réalisant tous les découpages possibles, en s'appuyant sur des tests binaires de la forme :

- $x_i \leq v$ pour les variables quantitatives, où v est une valeur prise par la variable x_i ;
- $x_i \in E$ pour les variables qualitatives, où $\{E, C_E\}$ est une partition à deux classes de la variable (en effet, il y a autant de tests possibles que de partitions à deux classes de la variable à modalités).

Soit $\mathbb{X} = \{\mathbb{X}_1, \mathbb{X}_2\}$ un découpage de la base initiale. En reprenant les notations utilisées en introduction, l'entropie totale de la base est donnée par :

$$H(\mathbb{X}) = -\sum_{k=1}^{K} \varpi_k . log_2(\varpi_k) \tag{4.5}$$

Où ϖ_k représente la proportion de la classe C_k dans la base de données. Posons par ailleurs $q_1 = 1 - q_2 = \frac{card(\mathbb{X}_1)}{card(\mathbb{X})}$, l'entropie après découpage de la base initiale devient :

$$q_1 . H(\mathbb{X}_1) + q_2 . H(\mathbb{X}_2) \tag{4.6}$$

Le gain d'information induit par ce découpage est alors simplement la différence entre les deux quantités :

$$G(\mathbb{X}) = H(\mathbb{X}) - \sum_{i=1}^{2} q_i . H(\mathbb{X}_i) \tag{4.7}$$

La division optimale de la base est celle qui maximise le gain d'information. La construction des nœuds de l'arbre est donc réalisée de manière récursive. Un nœud est considéré terminal si toutes les observations qu'il contient appartiennent à la même classe.

Élagage de l'arbre

Pour répondre à la deuxième question, nous reprenons l'exemple illustratif portant sur la classification des assurés sur la base des attributs âge et ancienneté. Le développement de l'arbre maximal permettant de s'ajuster complètement aux données d'apprentissage est donné en figure 4.6.

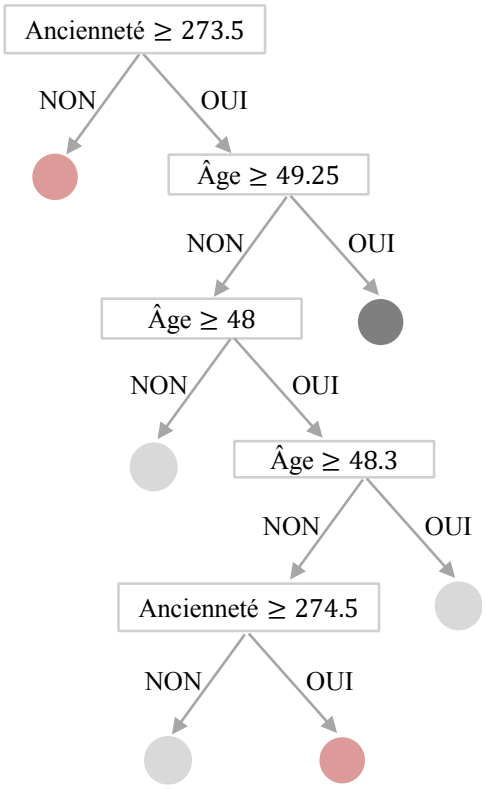


Figure 4.6. Développement maximal de l'arbre de décision : illustration du sur-apprentissage

Force est de constater que l'arbre construit est peu exploitable car il est trop ajusté aux données d'apprentissage. Pour rappel, l'objectif final est de généraliser la classification à de nouveaux assurés qui ne figurent pas initialement dans notre base d'apprentissage.

Pour remédier à cette problématique de sur-apprentissage, nous avons recours à l'élagage de l'arbre. En d'autres termes, nous avons besoin d'un critère nous permettant d'arrêter le développement de l'arbre à la profondeur optimale.

Plusieurs méthodes d'élagage ou de *pruning* existent dans la littérature. Une première méthode heuristique consiste à limiter le développement de l'arbre pendant sa construction : le développement d'un nœud s'arrête quand il s'agit de construire des feuilles n'améliorant pas considérablement le pouvoir de prédiction. Cette méthode nous permet par la même occasion de gagner en temps de calcul car cela nous évite de faire le développement complet de l'arbre. En revanche, cette méthode conduit dans certains cas au sous-apprentissage. En

d'autres termes, si une étape η de développement de l'arbre n'améliore pas significativement son pouvoir de prédiction, il se peut que le développement à l'étape $\eta + 1$ soit bénéfique. De ce fait, limiter le développement à l'étape η sans vérifier les étapes suivantes n'est pas forcément la solution idéale.

Dans le cadre de nos travaux, nous aurons recours à une autre méthode d'élagage. En effet, l'idée consistera à créer les arbres de décision avec les différentes profondeurs possibles, allant jusqu'à la profondeur maximale permettant de s'ajuster complètement aux données d'apprentissage. Ensuite, pour chaque arbre construit, nous mesurons sa précision grâce à la méthode de validation croisée évoquée précédemment. La profondeur retenue sera la profondeur minimale, au-delà de laquelle la précision ne s'améliore pas considérablement.

4.2.4 Limites des arbres de décision et apprentissage ensembliste

Si les arbres de décisions présentent beaucoup d'avantages tels que l'aisance d'interprétation, la rapidité d'exécution ou encore le fait que cela fonctionne sans émettre d'hypothèse particulière sur les données, il n'en demeure pas moins qu'ils présentent quelques limitations. La limitation principale des arbres de décision est l'instabilité.

En effet, l'arbre obtenu par l'algorithme décrit précédemment est très sensible aux données d'apprentissage, en d'autres termes, un changement même très marginal dans les observations ou dans les variables considérées peut conduire à la construction d'un arbre de décision complètement différent, impactant ainsi la précision de l'algorithme lorsqu'il s'agit de prédire de nouvelles observations.

Des travaux de recherche notables ont permis de traiter le problème d'instabilité des arbres de décision. Les méthodes couramment utilisées sont dites d'apprentissage ensembliste ou *Ensemble learning*. L'idée principale de ces méthodes est de combiner des classificateurs faibles [1] afin de construire un modèle plus performant. Dans notre cas, les classificateurs faibles correspondront simplement à des arbres de décision.

Nous aborderons deux familles de modèles d'apprentissage ensembliste : la première repose sur des stratégies aléatoires (*Random Forests*, introduite par [Breiman 2001]), la deuxième sur des stratégies adaptatives (*Adaptative Boosting*, introduite par [Freund 1996]).

4.2.5 Les forêts aléatoires

Dans le but de réduire l'instabilité des arbres de décision individuels, l'idée est de construire une forêt aléatoire, c'est-à-dire plusieurs arbres de décision sur la base d'échantillons *Bootstrap* issus de la base d'apprentissage. Chaque arbre contribue alors à la prédic-

1. Pour une classification binaire d'observations uniformément distribuées, un classificateur faible permet de déterminer correctement la classe d'appartenance d'une observation avec une probabilité $\geq 50\%$.

tion finale par le biais d'un vote majoritaire. En d'autres termes, la classe d'appartenance d'un nouvel individu correspond à celle ayant obtenue le nombre maximal de votes par l'ensemble des arbres de décisions créés.

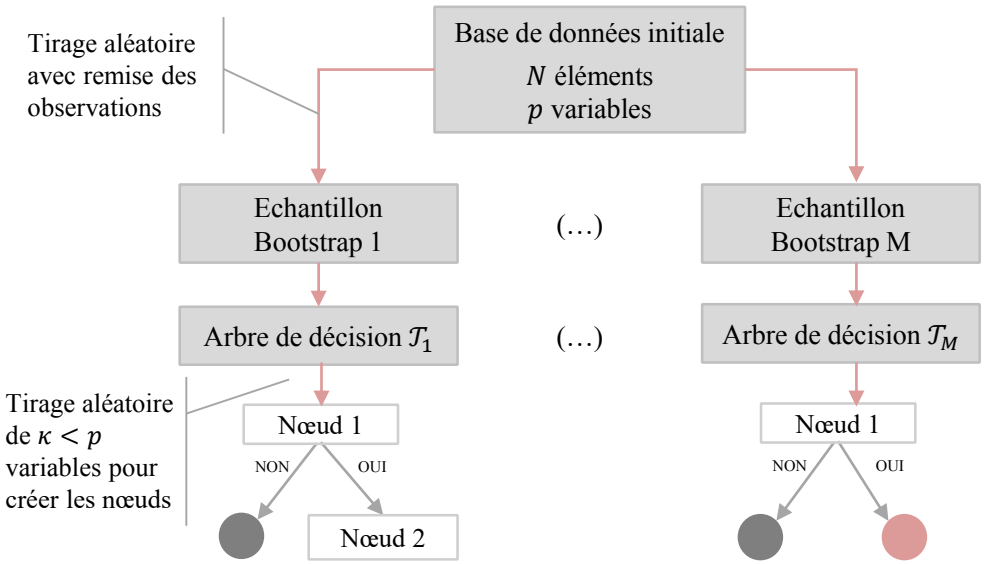


Schéma 4.3. Illustration de la méthode des Random Forests

Le caractère aléatoire de la méthode est lié à deux éléments :

- D'abord, le tirage avec remise des observations de la base initiale afin de construire des sous-ensembles d'observations ;
- Ensuite, le choix des variables explicatives à prendre en compte lors du découpage des zones de l'espace (i.e. création des nœuds de l'arbre) se fait aussi aléatoirement selon un nombre $\kappa < p$ de variables.

L'objectif de ces tirages aléatoires est d'introduire plus de variabilité dans les classificateurs créés. Certes, les arbres de décisions construits sur la base d'un sous-ensemble d'individus et d'un nombre restreint de variables seront moins performants individuellement que l'arbre de décision complet. En revanche, le résultat final est plus précis car la prédiction d'un nouvel individu se fera sur la base d'un vote majoritaire impliquant l'ensemble des classificateurs construits.

Les paramètres M et κ dépendent du problème considéré. La procédure consiste à augmenter la valeur M jusqu'à la stabilisation du score de prédiction. Le choix du paramètre κ se fait également en testant plusieurs valeurs, en général $\kappa = \sqrt{p}$ conduit à des résultats satisfaisants en classification.

4.2.6 Le Boosting adaptatif

Le principe des méthodes de Boosting est proche de celui utilisé pour les Random Forests, dans le sens où cette famille de méthodes consiste à s'appuyer sur plusieurs classificateurs faibles afin de réduire la variance. Cependant, contrairement à la méthode des Random Forests, les méthodes de Boosting permettent de réduire le biais (i.e. l'erreur liée au sous-apprentissage).

Supposons que notre base d'apprentissage totale contient très peu d'assurés pour une classe donnée, alors l'algorithme d'apprentissage aura un pouvoir de prédiction faible pour cette classe. La réalisation de tirages aléatoires pour construire des échantillons Bootstrap à partir de la base complète ne résout pas ce problème. En effet, l'algorithme final sera certes plus stable quant aux variations des données, mais il sera aussi biaisé que l'arbre de décision lorsqu'il s'agit de prédire des observations appartenant à la classe minoritaire considérée : le Boosting permet de répondre à cette problématique.

Nous avons vu que les forêts aléatoires s'appuient sur la construction **parallèle** de plusieurs arbres de décision. Pour le Boosting, il s'agit plutôt d'une construction **séquentielle**.

Le Boosting adaptatif consiste à initialiser l'ensemble des individus avec un poids identique. Ensuite, pour chaque étape de l'algorithme itératif, les individus mal classés se voient attribuer un poids plus important, alors que le poids des individus bien classés reste inchangé. Cela signifie que le modèle focalise ses efforts sur les observations difficiles à classer.

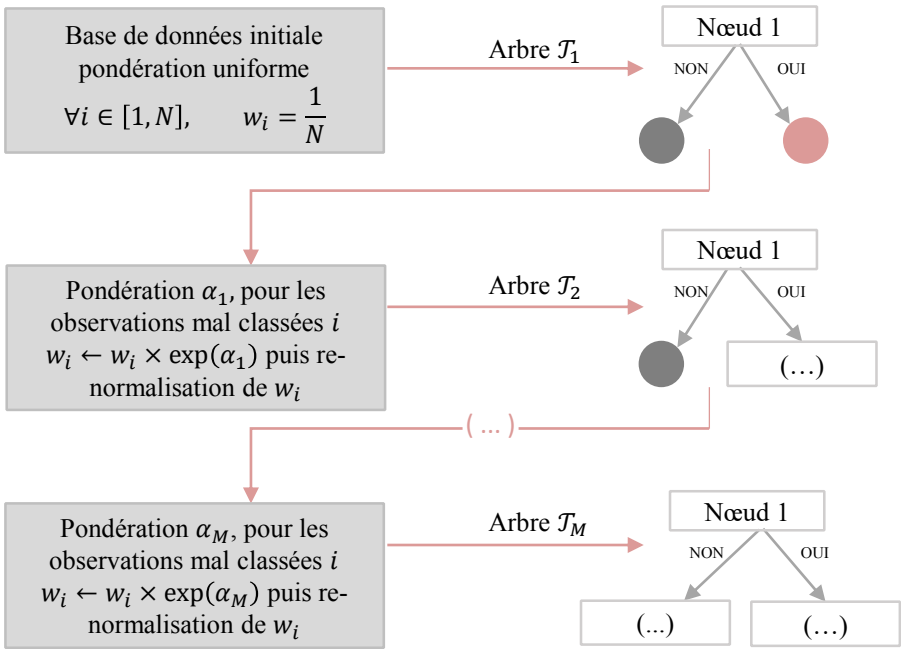


Schéma 4.4. Illustration de la méthode du Boosting adaptatif

L'application des poids sur les individus mal classés peut se faire soit en utilisant le ré-échantillonage des observations ou également lors du calcul de l'entropie. En guise d'exemple, on considère les observations $\{a, b, b\}$ et les poids associés $\{\frac{2}{4}, \frac{1}{4}, \frac{1}{4}\}$. L'entropie sans prise en compte des poids est donnée par $-\frac{1}{3}log_2(\frac{1}{3}) - \frac{2}{3}log_2(\frac{2}{3}) = 92\%$. En tenant compte des poids de chaque observation, l'entropie devient : $-\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{4}) = 100\%$. L'application des poids a ainsi rendu uniforme la distribution des classes.

L'algorithme de Boosting utilisé dans le cadre de notre étude est une généralisation du modèle AdaBoost ([Freund 1996]) à des problèmes de classification en classes multiples. L'algorithme détaillé ainsi que le mode de calcul des coefficients $\alpha_m, m \in [1, M]$ sont fournis en annexe, mais cela peut être présenté par le schéma 4.4.

Pour une observation donnée, la classe d'appartenance fournie par le modèle correspond alors à un vote pondéré des arbres $(\mathcal{T}_1, ..., \mathcal{T}_M)$. M correspond au nombre d'itérations, paramètre défini en entrée de l'algorithme.

Ce modèle s'avère très efficace lorsque la base de données manipulée présente une forte asymétrie au sein des classes. Nous allons voir qu'il permet d'améliorer très légèrement la qualité de prédiction lors de son application sur les bases de données que nous allons manipuler compte-tenu de leur caractère suffisamment équilibré.

Chapitre 5

Agrégation du portefeuille

5.1 La méthodologie d'agrégation

Nous avons maintenant défini et analysé l'ensemble des éléments entrant en jeu dans la méthodologie d'agrégation développée dans le cadre de ce mémoire, notamment les données relatives aux assurés (MPP), les hypothèses ainsi que le modèle de projection.

Pour commencer, il convient de mentionner que le principe de la méthode d'agrégation proposée repose sur le constat suivant :

Deux assurés générant des cashflows identiques ne sont pas forcément proches en termes de caractéristiques sous-jacentes.

Afin d'illustrer ce constat, prenons le cas d'une garantie temporaire décès à horizon 1 an. Sur la base des tables de mortalité du moment TH/TF-002, nous avons :

Assuré	Probabilité de décès dans l'année
Femme d'âge 46 ans	$0,1951\%$
Homme d'âge 38 ans	$0,1953\%$

Table 5.1. Illustration du critère de similarité entre deux assurés

Les deux individus en portefeuille, génèrent quasiment les mêmes flux futurs probables car les probabilités de décès sont très proches : il est donc possible de les agréger au sein du même Model Point sans perte significative de précision dans les calculs. Cependant, l'application de l'agrégation aux caractéristiques intrinsèques des assurés ne peut pas conduire à la classification des deux individus au sein du même groupe, compte-tenu de leur différence d'âge et de sexe.

$\Longrightarrow$ Le réflexe naturel face à ce constat est donc de s'appuyer sur les cashflows futurs probables lors de la construction des groupes homogènes de Model Points, au lieu d'appliquer l'agrégation aux caractéristiques des assurés. Cependant, cela met en lumière une problématique relative à la pérennité de la méthode :

Faut-il réitérer l'évaluation des cashflows futurs probables à chaque fois que l'on souhaite appliquer la méthode d'agrégation sur le portefeuille de passif vieilli ou sur un nouveau portefeuille ?

Pour répondre à cette problématique, nous distinguons deux cas de figure :

– **Cas 1 :** quand la simulation d'une base exhaustive contenant l'ensemble des combinaisons de caractéristiques possibles (d'âge, d'ancienneté, de sexe etc.) est réalisable et que les ressources informatiques à disposition permettent de manipuler les flux futurs associés à cette base d'individus. Ces flux peuvent être exploités pour construire des groupes représentatifs qui pourront par extension représenter le portefeuille réel. Dans ce cas, un individu faisant partie d'un portefeuille quelconque appartient forcément à notre base exhaustive : la projection de la base exhaustive se fait donc une seule fois.

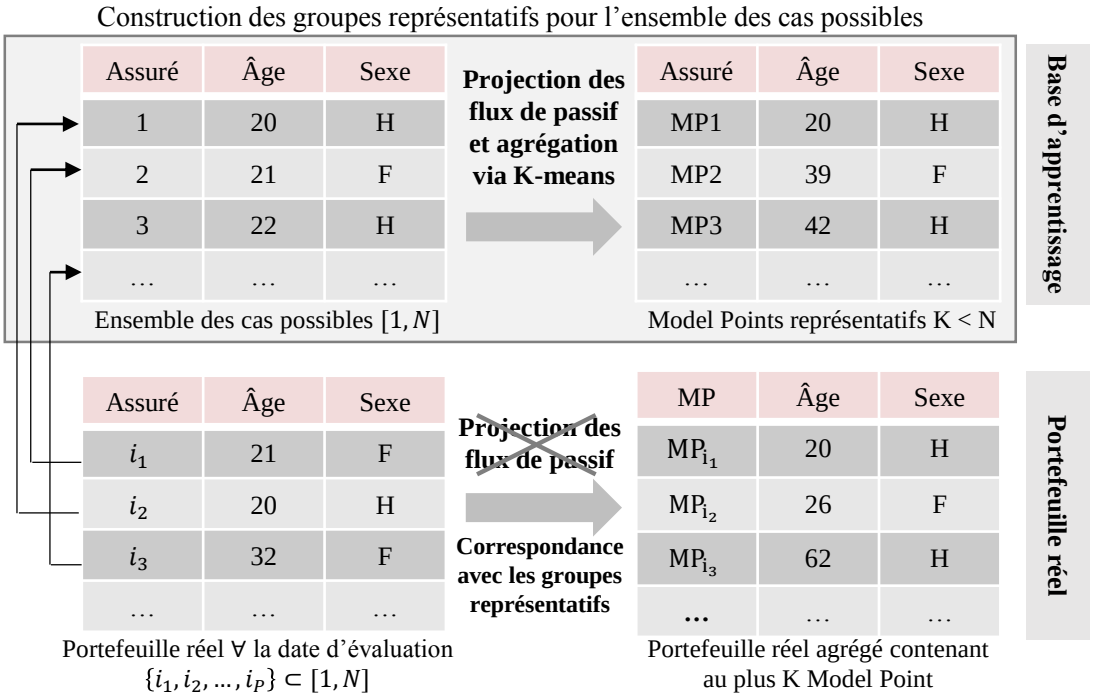


Schéma 5.1. Généralisation de l'approche d'agrégation - Cas 1

– **Cas 2 :** les ressources informatiques à disposition ne permettent pas d'exploiter les données issues de la base exhaustive à cause du nombre trop élevé de combinaisons

possibles.
Dans ce cas, les méthodes d'apprentissage automatisé fournissent une réponse très forte à cette problématique. L'idée consiste à construire un algorithme capable de de reconnaître des groupes similaires à l'exemple précédent portant sur les deux assurés, et ce en entraînant cet algorithme sur une base de données adéquate. Là encore, la projection des flux issus de cette base de données ne sera réalisée qu'une seule fois, car l'algorithme de classification supervisée aura appris à regrouper les individus issus de n'importe quel portefeuille.

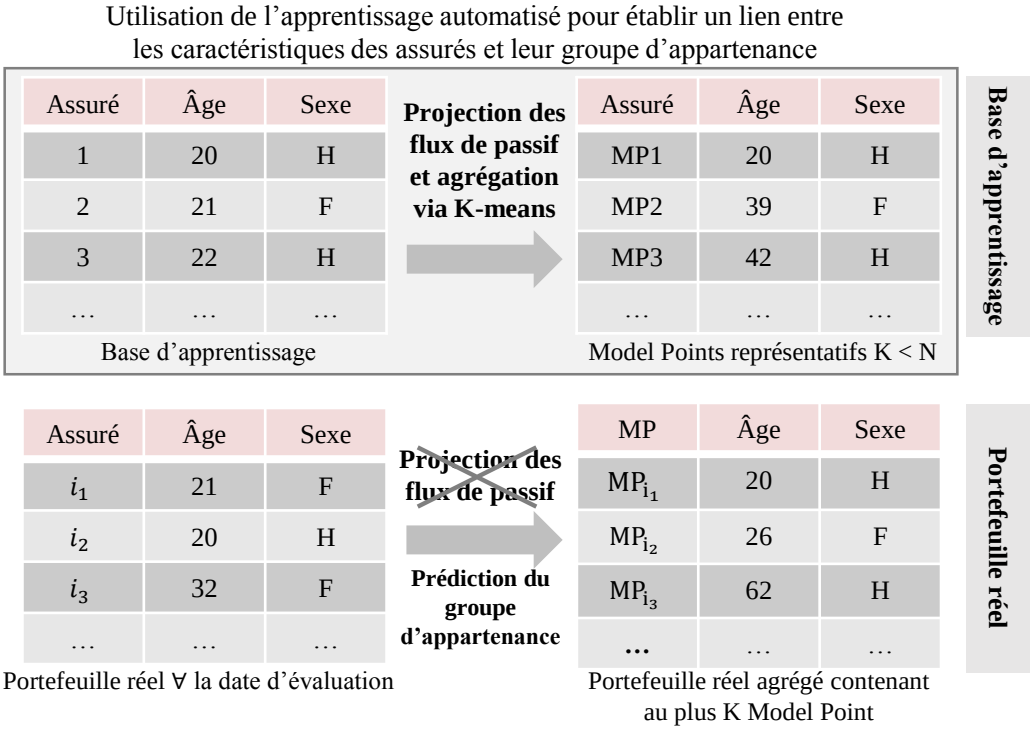


Schéma 5.2. Généralisation de l'approche d'agrégation - Cas 2

Si les méthodes de classification à utiliser et la nature du passif à agréger ont déjà été abordées dans les chapitres précédents, quelques questions sont à instruire : quels sont les cashflows futurs sur lesquels s'appuyer lors de la construction des groupes ? à quoi correspond le nombre optimal de groupes représentatifs à créer pour chaque garantie ? enfin, à quoi correspond la base d'apprentissage adéquate pour l'apprentissage automatisé ?

Les réponses à ces questions seront apportées dans les sections suivantes, en prenant en compte les garanties étudiées, l'approche de modélisation ainsi que la taille du portefeuille à disposition.

5.2 Déclinaison opérationnelle et définition des paramètres

L'objectif de cette partie est de détailler les différentes étapes permettant d'appliquer la méthode d'agrégation dont la théorie a été présentée précédemment. Par la même occasion, nous préciserons les cashflows à utiliser pour chaque garantie, ainsi que le processus de simulation de notre base d'apprentissage.

5.2.1 Étapes de mise en place

D'un point de vue opérationnel, la déclinaison de la méthode d'agrégation proposée s'effectue, par garantie, selon les étapes définies dans le diagramme 5.1. En résumé, les groupes représentatifs sont d'abord construits en s'appuyant sur une base d'apprentissage, ensuite l'algorithme de prédiction des groupes est développé en vue de son application aux données réelles.

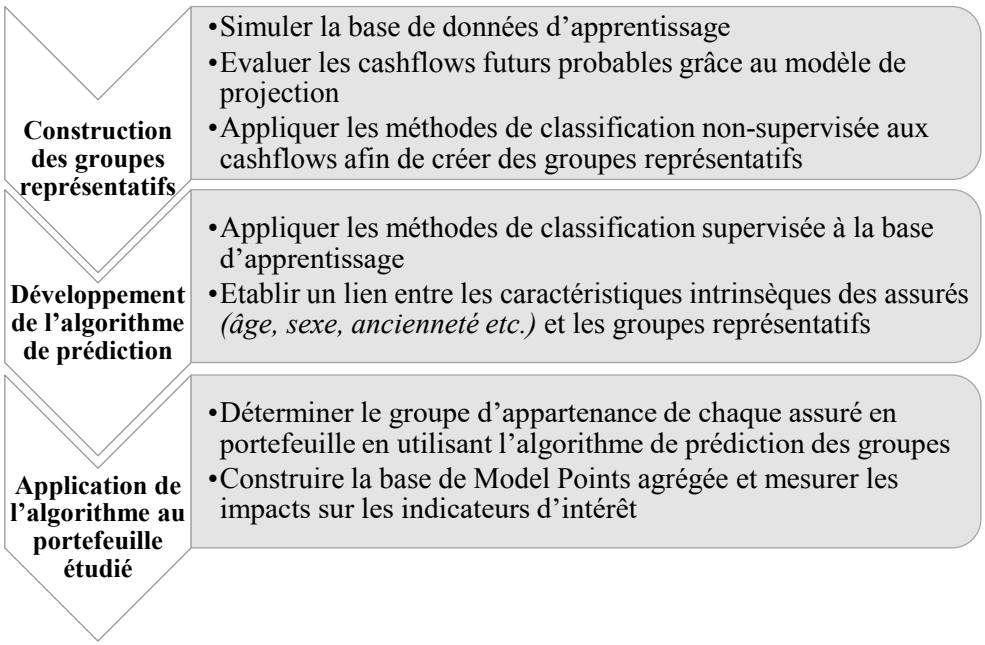


Diagramme 5.1. Étapes de mise en place de la méthode d'agrégation du passif

5.2.2 Choix des cashflows par garantie

Intéressons-nous à présent aux cashflows requis pour la première étape d'agrégation. Comme évoqué précédemment, les cashflows qui seront utilisés correspondent à des sorties

du modèle de projection, ils sont la résultante de l'approche de modélisation ainsi que des hypothèses retenues. Si le choix des cashflows dépend naturellement de la garantie étudiée, un critère important doit néanmoins être respecté quelle que soit la garantie : les hypothèses sous-jacentes doivent être suffisamment stables dans le temps, et ce dans le but d'assurer la pérennité de la méthode d'agrégation.

Or, les prestations futures probables répondent au critère de stabilité, car leur calcul repose sur des lois de probabilité (par exemple, les tables réglementaires de mortalité et les lois de maintien dans l'état d'incapacité ou d'invalidité). Par conséquent, la méthode d'agrégation construite sera aussi stable que les tables réglementaires retenues dans les calculs des prestations futures. Ceci dit, dans le cas d'un changement de table, la même méthode peut être utilisée pour agréger le passif en réévaluant la chronique des prestations suite au changement de loi.

En revanche, l'utilisation des prestations futures probables présente une limite pour certaines garanties. En particulier, pour les rentes d'éducation et les rentes de conjoint, le taux technique contractuel n'a pas d'impact sur le montant des prestations futures. Nous aurons dans ce cas recours aux provisions mathématiques : ces derniers sont non seulement sensibles aux taux techniques, mais elles reposent aussi sur des hypothèses aussi stables que les prestations.

L'arbitrage entre prestations futures probables ou la provision mathématique ne sera bien entendu pas généralisé sur l'ensemble des garanties pour la simple raison que pour certaines garanties, le taux technique n'est pas une hypothèse propre à la police d'assurance mais plutôt une hypothèse générale déterminée par les conditions financières. Par exemple, pour la garantie invalidité, le calcul de la provision mathématique repose sur le taux technique non-vie à la date d'évaluation et non sur un taux contractuel propre à chaque contrat.

Compte-tenu de l'analyse des différents éléments cités précédemment, les choix des cashflows d'agrégation par garantie sont récapitulés dans le tableau 5.2.

Garantie	Cashflow d'agrégation
Incapacité	Prestations futures probables
Invalidité	
Maintien de la garantie décès	
Rente d'éducation	Provisions mathématiques projetées
Rente de conjoint	

Table 5.2. Choix des cashflows utilisés pour l'agrégation

5.2.3 Simulation de la base de données d'apprentissage

Avant d'entamer l'agrégation du portefeuille à disposition, nous allons simuler notre base de données d'apprentissage. Deux étapes essentielles reposeront sur la simulation de cette base de données :

- D'une part, cette base est simulée dans la perspective d'identifier les Model Points candidats à représenter les assurés que nous avons au sein du portefeuille réel en appliquant les méthodes de classification non-supervisée ;
- D'autre part, la simulation de cette base nous permettra d'établir les liens qui existent entre les caractéristiques des assurés et leur groupe représentatif.

Le schéma suivant décrit la démarche retenue pour générer cette base d'apprentissage :

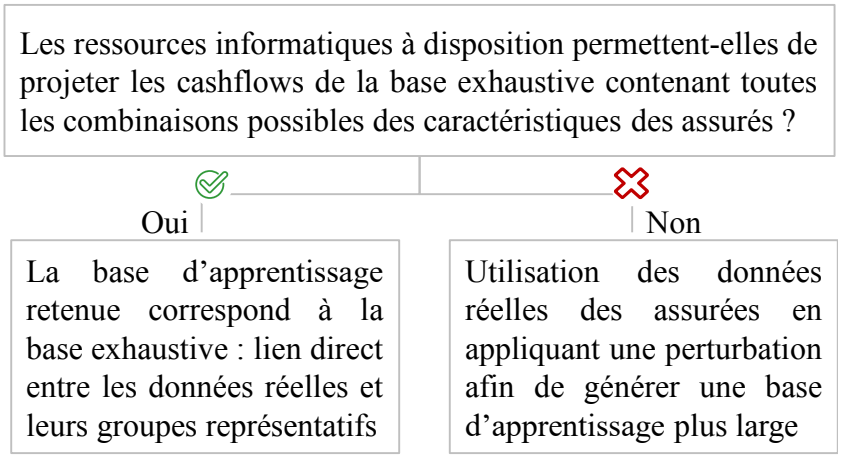


Schéma 5.3. Construction de la base de données d'apprentissage

A ce stade, rappelons un aspect de modélisation ayant un impact sur la taille des bases d'apprentissage : le pas de temps utilisé dans notre modèle de projection est mensuel. Ainsi, pour une année de naissance donnée, nous avons 12 cas possibles de Model Points correspondant à des mois de naissance différents.

Quand les ressources informatiques à disposition ne permettent pas d'exploiter les cashflows de la base exhaustive (i.e. contenant l'ensemble des cas possibles) à cause de sa taille trop élevée, la simulation de la base d'apprentissage est réalisée par tirage aléatoire, tout en veillant à respecter les deux contraintes suivantes :

1- La base simulée doit refléter les caractéristiques réelles des assurés
2- Les observations doivent être suffisamment équilibrées, afin d'éviter le problème de l'asymétrie des classes (cf. 4.2.1).

Pour les différentes raisons citées précédemment, nous partons du portefeuille réel et nous appliquons des perturbations aux différentes variables caractérisant les assurés, afin

de générer une base de données plus large. Ces perturbations reproduisent les différentes variations que peut subir le passif, notamment le vieillissement des assurés ou les changements d'état de sinistralité. Notre base d'apprentissage aura d'un côté l'avantage de refléter le portefeuille réel. D'un autre côté, cette méthode de simulation permet contrecarrer de problème d'asymétrie des classes.

5.2.4 Critère de validation de l'agrégation

La méthode des K-means sera appliquée aux cashflows issus de la projection des bases d'apprentissage. Comme mentionné précédemment, cette méthode prend en paramètre le nombre de classes à construire à partir de la base initiale. Il convient donc de définir un indicateur permettant de fixer le nombre de groupes représentatifs par garantie, tout en maîtrisant l'impact de l'agrégation.

A chaque pas de projection t, les cashflows du passif ont une incidence sur les comptes de participation aux bénéfices (car le calcul du solde technico-financier s'appuie entre autre sur les flux du passif) ainsi que sur l'interaction Actif-Passif (car le niveau de l'actif est ajusté à chaque pas de temps en fonction des flux du passif pour assurer l'équilibre du bilan). Dès lors, le critère de validation à retenir doit prendre en compte les impacts de l'agrégation du passif non seulement sur la valeur actualisée des flux futurs, mais également sur leur variation tout au long de la projection.

Dans le but d'introduire le critère de validation qui sera retenu, notons, pour un pas de temps t, CF_t la somme des cashflows d'un groupe d'assurés avant agrégation et CF_t^{ag} le montant de prestations obtenu après agrégation du groupe. La valeur absolue de l'erreur relative en t est alors donnée par :

$$Er(t) = \left| \frac{CF_t - CF_t^{ag}}{CF_t} \right| \tag{5.1}$$

Maîtriser l'impact de l'agrégation à chaque pas de temps revient à fixer un seuil sur la valeur maximale prise par l'erreur absolue. Par exemple, pour un taux d'erreur inférieur à 1% quel que soit le pas de temps, et en posant T l'horizon de projection il faut vérifier que :

$$\max_{t \in [1,T]} Er(t) < 1\% \tag{5.2}$$

Avec ce critère, nous pouvons affirmer que sur l'ensemble des garanties modélisées et quel que soit le pas de temps $t \in [1, T]$, les cashflows de prestations associés au portefeuille agrégé présentent une erreur relative inférieure à 1% par rapport à la projection du portefeuille initial non-agrégé. Il convient de mentionner à ce stade qu'en ayant recours à ce critère, nous observons généralement des taux d'erreur relative très faibles en début de projection (plus faibles que le seuil de 1%).

Nous verrons par la suite que l'écart final sur le bilan Solvabilité II est négligeable. En effet, la chronique des cashflows futurs du passif étant décroissante (car l'agrégation se fait sur le stock de contrats et non sur les affaires nouvelles), l'erreur relative peut être élevée en fin de projection sans que les montants en jeu soient matériels.

D'autre part, le Best Estimate étant obtenu en actualisant les flux futurs probables, l'impact de l'agrégation sur les cashflows en fin de projection sera fortement réduit par l'effet de l'actualisation.

5.2.5 Propriété d'additivité et représentation par le barycentre

Pour finir l'introduction de la méthode, il convient de mentionner que pour l'ensemble des garanties étudiées, nous nous appuyons sur le caractère additif des flux projetés. En effet, pour deux assurés i et j ayant exactement les mêmes caractéristiques (âge, sexe, ancienneté etc.) les prestations projetées sont parfaitement additives : notons r_i et r_j les montants respectifs des indemnités prévues contractuellement (rente ou forfait mensuel) pour les assurés i et j, nous avons :

$$CF_t(r_i, i) + CF_t(r_j, j) = CF_t(r_i + r_j, i) = CF_t(r_i + r_j, j) \tag{5.3}$$

Où $CF_t(r, i)$ représentent les prestations, au pas de temps t, dues à un assuré i ayant une indemnité r. Cela signifie que l'agrégation des individus ayant les mêmes caractéristiques n'induit pas de perte de précision dans les calculs.

Par extension et comme mentionné dans la section 4.1.3, nous représentons un groupe d'individus $\{i_1, ..., i_n\}$ par l'individu i_k le plus proche au barycentre du groupe de sorte que :

$$\overbrace{\sum_{l=1}^{n} CF_t(r_{i_l}, i_l)}^{\text{flux avant agrégation}} \simeq \overbrace{CF_t(\sum_{l=1}^{n} r_{i_l}, i_k)}^{\text{flux après agrégation}} \tag{5.4}$$

L'objectif est donc de maîtriser l'erreur induite par cette représentation sur la base du critère de validation fixé précédemment.

5.3 Application sur le portefeuille et analyse des résultats

L'objectif de cette partie est d'appliquer la méthode d'agrégation sur les données de l'entité étudiée. Nous présentons d'abord les impacts sur la chronique des provisions mathématiques projetées et sur les flux de prestations probables pour l'ensemble des garanties concernées, avant de faire une analyse de la vision Solvabilité II dans le chapitre suivant.

Invalidité

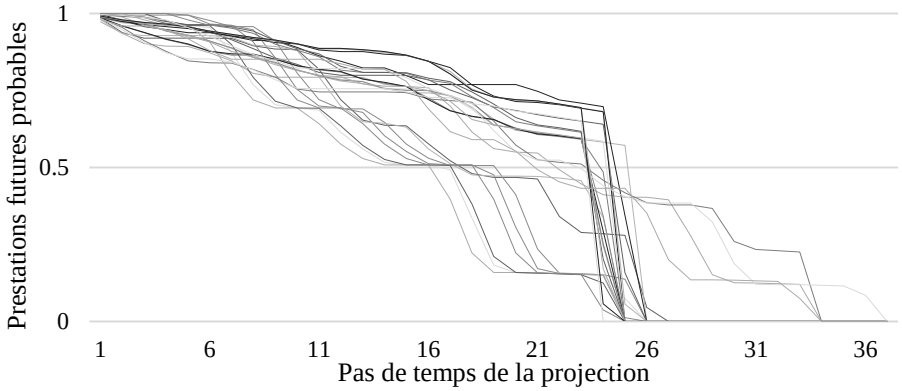
Nous nous intéressons d'abord à la garantie invalidité. Pour rappel, elle consiste à verser une indemnité aux assurés en état d'invalidité permanente. La projection des prestations futures probables en invalidité dépend de deux paramètres : l'âge de l'assuré et sa date d'entrée en invalidité. L'âge des assurés en invalidité est limité à 62 ans pour les contrats commercialisés par l'entité étudiée. L'âge des assurés est donc compris entre 18 et 62 ans, ce qui correspond à l'intervalle mensuel $[216, 744]$.

Pour un âge d'assuré k appartenant à l'intervalle $[216, 744]$, nous avons $k - 216 + 1$ possibilités pour la date d'entrée en invalidité. Ce qui nous permet d'écrire que le nombre total de cas possibles en invalidité est donné par :

$$\sum_{k=216}^{744} (k - 216 + 1) = 140185 \tag{5.5}$$

Les ressources informatiques à disposition ne permettent pas d'exploiter la base de données contenant les flux futurs relatifs à ce nombre très important d'individus. Nous allons donc réaliser une simulation d'une base d'apprentissage contenant 70 000 lignes, ce qui correspond globalement à la limite du nombre d'assurés que nous pouvons manipuler dans le cadre de notre étude, compte-tenu de la capacité des ressources informatiques utilisées.

La simulation de notre base d'apprentissage repose sur la perturbation des données réelles du portefeuille. Pour une ancienneté i, nous avons plusieurs âges possibles avec une moyenne μ_i et un écart-type σ_i. L'idée est donc de simuler aléatoirement plusieurs âges différents pour chaque ancienneté, selon une loi gaussienne $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ afin de construire une base de données plus large, ayant les mêmes caractéristiques que le portefeuille initial.



Graphique 5.1. Illustration de la projection des prestations futures relatives aux individus de la base d'apprentissage en invalidité (échantillon de 26 individus)

Les flux du passif de la base d'apprentissage sont évalués grâce au modèle de projection

passif seul. Nous obtenons alors les chroniques des prestations futures, avec une indemnité mensuelle unitaire pour l'ensemble des individus de la base d'apprentissage (cf. figure 5.1).

Ensuite, nous appliquons la méthode de classification non-supervisée retenue sur les cashflows projetés afin de créer les groupes représentatifs pour la garantie étudiée. La démarche consiste à faire varier le nombre de groupes à former, de manière à construire une agrégation respectant le critère fixé (erreur relative maximum inférieure à 1%).

Pour chaque stratégie d'agrégation réalisée, les flux de prestations sont réévalués par le modèle de projection afin de les comparer aux flux avant agrégation, puis de calculer l'écart relatif à chaque pas de temps suite aux regroupements des assurés. Rappelons que pour chaque groupe identifié par la méthode des K-means, nous prenons comme individu représentatif celui dont la distance est la plus proche au barycentre du groupe.

Les résultats sont présentés dans le tableau 5.3 (pour rappel, la base d'apprentissage contient initialement 70 000 individus). L'agrégation de la base avec 500 groupes fournit des résultats très satisfaisants. En effet, l'erreur relative est inférieure à 0.85% quel que soit le pas de temps. En plus, l'erreur observée en début de projection est faible.

Nombre de groupes	Erreur relative maximale	Erreur relative moyenne
250	2.00%	0.26%
500	0.85%	0.12%
750	0.58%	0.10%

Table 5.3. Application des K-means sur la base d'apprentissage en invalidité

Nous pouvons donc représenter n'importe quel portefeuille d'assurés en invalidité par une nouvelle base agrégée contenant au maximum 500 Model Points avec une perte de précision inférieure à 1% sur les flux du passif projetés.

Cependant, pour appliquer cette agrégation sur notre portefeuille réel, nous devons disposer des règles permettant de faire le lien entre les caractéristiques des assurés (en l'occurrence leur âge et leur date d'entrée en invalidité) et le groupe auquel ils appartiennent parmi les 500 groupes représentatifs.

L'étape suivante consiste à appliquer les algorithmes d'apprentissage supervisé, sur les caractéristiques des individus présents dans la base d'apprentissage et leur groupe d'appartenance. Il convient donc de se donner un moyen permettant de vérifier la validité des prédictions fournies par l'algorithme. Nous utilisons la méthode de la validation croisée des L-folds présentée précédemment pour valider notre algorithme de prédiction. Nous fixons le paramètre L à 4 (nous aurons ainsi des bases de validation contenant 25% des individus).

La mesure de précision qui sera retenue à ce titre est le pourcentage moyen d'individus,

dans la base de validation, pour lesquels l'algorithme a pu prédire correctement le groupe d'appartenance. Les résultats obtenus pour les différentes méthodes sont récapitulés dans le tableau ci-dessous.

Méthode	Score moyen
Arbres de décision	96.60%
Random Forests	97.03%
Adaptive Boosting	97.56%

Table 5.4. Application des méthodes de classification supervisée en invalidité

Les trois méthodes fournissent des résultats satisfaisants. Nous retenons la méthode du Boosting adaptatif, qui permet d'améliorer le score en mettant plus de poids aux individus mal classés. Pour les observations potentiellement mal classées, le traitement le plus simple consisterait à ne pas les inclure dans l'agrégation. Nous pouvons les identifier en fixant un seuil (pour la probabilité d'appartenance à un groupe) au-dessous duquel les assurés ne seront pas intégrés au sein d'un groupe représentatif mais plutôt projetés séparément.

Dans le cadre de notre étude, nous faisons le choix d'intégrer l'ensemble des assurés dans le groupe représentatif suggéré par notre algorithme d'apprentissage. En effet, l'analyse des erreurs de prédiction montre que la grande majorité des individus mal classés sont à équidistance par rapport aux barycentres du groupe prévu initialement et de celui du groupe prédit par l'algorithme d'apprentissage. Ce résultat n'est pas étonnant car la propriété d'appartenance d'un individu à un groupe n'est pas exclusive. On considère l'exemple simplifié d'une agrégation s'appuyant sur deux flux (CF_1, CF_2) et pouvant être représentée sur un plan dans $\mathbb{R}^2$, l'illustration des individus conduisant à une erreur de classification est donnée dans la figure 5.1.

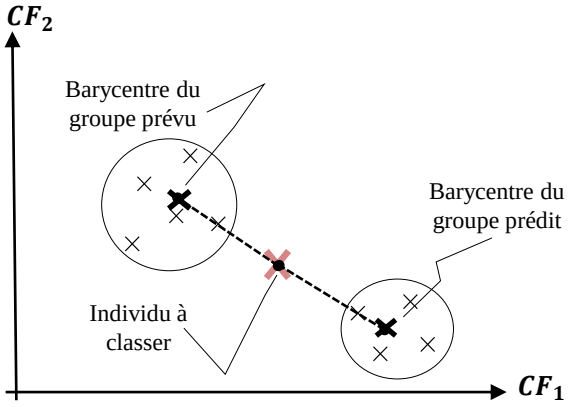
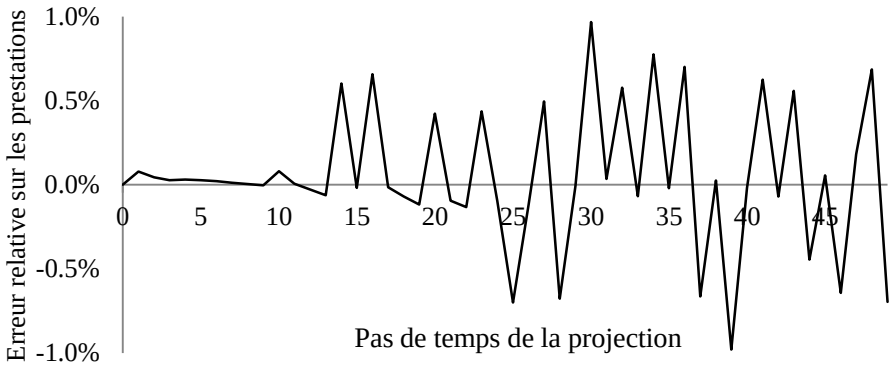


Figure 5.1. Analyse de l'erreur de prédiction *(l'individu à classer se situe à équidistance des barycentres du groupe prévu par les K-means et de celui prédit par l'apprentissage automatisé)*

En outre, le critère final de validation de l'algorithme d'apprentissage correspond plutôt aux résultats obtenus lors de son application sur notre portefeuille réel. Il convient donc d'appliquer notre algorithme au portefeuille réel, contenant initialement 19 443 assurés en invalidité. Le nombre de groupes obtenus après agrégation est de 408 avec un écart très faible sur les prestations.

Le graphe 5.2 présente les variations de l'erreur relative sur les prestations projetées en fonction du pas de temps de la projection : force est de constater que l'erreur relative est très proche de 0% en début de projection, elle augmente légèrement en fin de projection sans dépasser le seuil fixé de 1%.



Graphique 5.2. Variation de l'erreur relative sur les prestations suite à l'agrégation du portefeuille réel en invalidité

Cela aura donc peu d'impact sur les indicateurs Solvabilité II car d'une part, comme il s'agit d'une projection en run-off, le montant des prestations diminue avec le temps. D'autre part, le Best Estimate étant obtenu par actualisation des flux futurs, les écarts observés en fin de projection seront amortis par l'effet actualisation. Le calcul des provisions mathématiques sur la base des flux agrégés montre que l'erreur relative est de l'ordre de 0.13%.

À noter que les provisions pour sinistres à payer (PSAP) ne sont pas impactées par l'agrégation car elles sont calculées avec une hypothèse moyenne de cadences d'écoulement.

	Provisions totales	dont PM	dont PSAP
Invalidité	806 218	718 063	88 154
Erreur relative	**0.11%**	**0.13%**	**0.00%**

Table 5.5. Impact de l'agrégation sur les provisions en invalidité

Incapacité

Pour la garantie incapacité, la projection dépend également de la date de naissance et de la date d'entrée dans l'état d'incapacité. Cependant, une différence existe entre les deux garanties : l'ancienneté maximale en incapacité est limitée à 3 ans. Cela conduit bien évidemment à une base de données d'apprentissage moins volumineuse que celle construite pour l'invalidité.

En effet, pour tout âge compris entre 18 ans (216 mois) et 62 ans (744 mois), nous avons 37 anciennetés possibles, sauf pour les âges inférieurs à 21 ans dont l'ancienneté est forcément inférieure à 3 ans. Cela nous conduit au nombre suivant de possibilités :

$$\sum_{k=216}^{251} (k - 216 + 1) + \sum_{k=252}^{744} 37 = 18907 \tag{5.6}$$

L'outil de projection permet de projeter l'ensemble des combinaisons (âge, ancienneté), il est donc possible d'agréger n'importe quel portefeuille sans faire appel à l'apprentissage automatisé. Il suffit d'appliquer notre méthode de classification non-supervisée (K-means) à la base exhaustive, et de faire une correspondance entre les couples d'âge et d'ancienneté (i, j) et leurs Model Points représentatifs $MP_{(i,j)}$.

Nous réalisons donc la projection de la base exhaustive d'assurés afin d'en déduire les cashflows de prestations, sur lesquelles nous appliquons les K-means afin d'identifier le nombre de groupes représentatifs répondant au critère de validation fixé initialement.

Nombre de groupes	Erreur relative maximale	Erreur relative moyenne
500	3.89%	0.45%
1000	1.6%	0.29%
3000	0.58%	0.09%

Table 5.6. Application des K-means sur la base exhaustive en incapacité

Le nombre de groupes à créer, permettant de valider le critère d'erreur relative maximum inférieure à 1% est de 3000. Ce nombre est beaucoup plus élevé que celui des groupes représentatifs en invalidité, principalement pour deux raisons :

- D'une part, si le fait de construire les groupes à partir de la base exhaustive permet d'éviter l'erreur induite par les méthodes d'apprentissage automatisé, il n'en reste pas moins que cela conduise à créer des groupes qui ne seront pas forcément exploités lors de l'extension au portefeuille réel. Le nombre de groupes représentatifs créés lors de l'application des K-means est mécaniquement élevé car il permet de couvrir tous les cas possibles.

- D'autre part, un individu en état d'incapacité a trois transitions possibles (revenir à l'état valide, passer en état d'invalidité ou décéder) alors qu'un individu en état d'invalidité a une seule transition possible (cf. graphe sur la modélisation Markovienne). De ce fait, il y a plus de variabilité au niveau des prestations futures probables entre les assurés et ce même pour des âges et des anciennetés assez proches. Cela rend donc la segmentation plus complexe et conduit à créer un nombre plus important de groupes sous contrainte de minimisation de l'erreur.

Nous retiendrons la représentation des assurés avec 3000 Model Points : ce nombre correspond à la plus grande base de données en incapacité que l'on peut avoir, quel que soit le portefeuille à projeter, tout en ayant une erreur relative maximale inférieure à 0.58%.

L'étape suivante consiste à faire la correspondance entre les caractéristiques des assurés présents dans le portefeuille réel et leurs groupes représentatifs, car le portefeuille réel est simplement un sous-ensemble de la base exhaustive. Notre portefeuille réel contient initialement 16 419 assurés. La correspondance avec la base complète montre qu'ils peuvent être représentés par 2 157 groupes. Dans le tableau 5.7, nous examinons l'erreur d'agrégation obtenue sur les provisions d'inventaire car cela nous donne désormais une idée sur l'erreur qui sera observée sur le Best Estimate en incapacité.

	Provisions totales	dont PM	dont PSAP
Incapacité	448 576	350 283	98 293
Erreur relative	**0.31%**	**0.40%**	**0.00%**

Table 5.7. Impact de l'agrégation sur les provisions en incapacité

Maintien de la garantie décès

Pour cette garantie, nous avons exactement les mêmes paramètres à prendre en compte dans le dénombrement du nombre de cas possibles, à savoir la date de naissance et la date d'entrée dans l'état. Toutefois, l'état initial peut correspondre à celui de l'invalidité ou de l'incapacité. De ce fait, la dimension de la base d'apprentissage correspond simplement à la somme des dimensions des bases en incapacité et en invalidité, c'est-à-dire 159 092 combinaisons possibles.

Compte-tenu de la dimension de la base exhaustive, il convient de traiter séparément le maintien de la garantie décès en invalidité et en incapacité.

Maintien de la garantie décès - assurés dans l'état initial d'invalidité

Pour cette première catégorie d'assurés en portefeuille, nous procédons exactement de la même manière que le traitement de la garantie invalidité : nous construisons notre base d'apprentissage en appliquant des perturbations sur les caractéristiques des assurés présents dans le portefeuille réel.

Notre base d'apprentissage contient 70 000 individus, avec une distribution d'âge et d'ancienneté similaire à la base réelle, avec toutefois des perturbations reflétant les changements que l'on peut avoir sur notre portefeuille dans l'avenir.

Pour cette garantie, le versement du capital se fait au décès de l'assuré. Ceci dit, nous avons les mêmes transitions possibles que dans la garantie invalidité, à savoir le décès ou le maintien dans l'état en fonction de l'ancienneté. Cela nous conduit donc au même nombre de groupes représentatifs qui est de 500. L'application des K-means montre que cette représentation répond bien au critère de validation fixé (cf. tableau 5.8).

Nombre de groupes	Erreur relative maximale	Erreur relative moyenne
500	0.85%	0.17%

Table 5.8. Application des K-means sur la base d'apprentissage en maintien de la garantie décès - état initial : invalidité

De la même manière, l'étape suivante consiste à appliquer les trois méthodes de classification supervisée afin d'établir un lien entre les caractéristiques des assurés et leur groupe représentatif, et par extension avec l'individu le plus proche du barycentre du groupe. Les scores moyens obtenus, sur la base de la méthode de la validation croisée 4-folds, sont autour de 98%.

Méthode	Score moyen
Arbres de décision	95.70%
Random Forests	97.74%
Adaptive Boosting	98.40%

Table 5.9. Application des méthodes de classification supervisée - maintien de la garantie décès en invalidité

Maintien de la garantie décès - assurés dans l'état initial d'incapacité

Sur la garantie incapacité, nous avons vu que l'exploitation de la base exhaustive est réalisable, et que la construction de 3000 Model Points représentatifs répond bien au cri-

tère de validation de l'agrégation. Pour le maintien de la garantie décès en incapacité, les transitions des assurés sont identiques certes, mais c'est une garantie binaire : le capital est versé en cas de décès, les autres changements d'état n'ayant pas d'impact sur la garantie. De ce fait, il est possible de diminuer le nombre de groupes représentatifs sans que l'erreur dépasse le seuil fixé initialement. Pour cette garantie, nous retiendrons une représentation avec uniquement 2500 groupes représentatifs, cela permet de respecter le critère de validation comme le montre le tableau ci-dessous.

Nombre de groupes	Erreur relative maximale	Erreur relative moyenne
3000	0.47%	0.08%
2500	0.56%	0.12%

Table 5.10. Application des K-means sur la base d'apprentissage en maintien de la garantie décès - état initial : incapacité

Finalement, la correspondance entre les assurés dans le portefeuille réel et les groupes créés est réalisée : le portefeuille contient initialement 19 706 individus qui seront représentés uniquement par 1 784 Model Point. L'impact sur provisions mathématiques est de 0.57%, ce taux n'augmente pas dans la projection car l'erreur sur les prestations est limitée à 1% tout au long de la projection.

	Provisions totales	dont PM	dont PSAP
MGDC	156 589	135 515	21 074
Erreur relative	**0.50%**	**0.57%**	**0.00%**

Table 5.11. Impact de l'agrégation sur les provisions en maintien de la garantie décès

Rente d'éducation

Plusieurs paramètres entrent en jeu dans la projection des prestations futures en rente d'éducation. Le tableau 5.12 montre que le nombre de combinaisons possibles est de 56 448 compte-tenu des paramètres pris en compte dans la modélisation de cette garantie.

Pour cette garantie, nous n'aurons pas recours aux méthodes d'apprentissage automatisé car il est possible d'exploiter la base de données contenant les provisions mathématiques projetées, associées aux 56 448 combinaisons possibles des caractéristiques des assurés.

Dans un premier temps, nous faisons une projection passif seul de la base exhaustive des assurés, avec un montant unitaire pour la rente d'éducation, afin de calculer les provisions mathématiques projetées à l'horizon $T = 50$ ans pour chaque individu. Comme précisé

Variable	Nombre de possibilités
Âge strictement inférieur à 28 ans	28 x 12 = 336
Sexe	2
Catégorie socio-professionnelle	2 *(Parent Cadre / Non cadre)*
Taux technique	21 *(de 0% à 5% avec un pas de 0,25%)*
Fréquence de la rente	2 *(Mensuelle/Trimestrielle)*
Nombre total de possibilités	**56 448**

Table 5.12. Dimension de la base exhaustive en rente d'éducation

précédemment, la classification non-supervisée sera appliquée aux provisions et aux les prestations car ces dernières ne tiennent pas compte du taux technique de la rente.

Un constat intéressant émerge lors de l'application des K-means sur cette garantie : avec le même critère de validation, le nombre de groupes représentatifs à créer est très faible par rapport aux autres garanties étudiées jusqu'à présent : cela s'explique par les caractéristiques des individus impactant la modélisation. En effet, pour les garanties précédentes, seuls l'âge et l'ancienneté impactent la projection, pour la garantie rente d'éducation nous avons cinq variables d'intérêt. De ce fait, plusieurs combinaisons différentes des caractéristiques présentées dans le tableau ci-dessus conduisent à des chroniques de provisions assez proches au sens de la norme Euclidienne, ce qui permet in fine de réduire le nombre de groupes construits.

Les résultats de l'agrégation de la base exhaustive sont présentés dans le tableau 5.13 : il est possible de projeter n'importe quel portefeuille de rente d'éducation avec seulement 250 groupes, et une erreur relative inférieure à 1% sur les cashflows du passif projeté.

Nombre de groupes	Erreur relative maximale	Erreur relative moyenne
500	0.19%	0.07%
250	0.68%	0.15%

Table 5.13. Application des K-means sur la base d'apprentissage en rente d'éducation

Comme évoqué dans l'introduction de la méthode, les groupes créés n'incluent pas uniquement les individus dont l'âge est proche avec des caractéristiques (taux technique, fréquence de la rente etc.) identiques. En effet, nous pouvons trouver des profils très variés au sein du même groupe, les individus sont agrégés parce que leurs chroniques de provisions mathématiques sont proches au sens de la mesure de similarité utilisée dans les K-means.

Pour s'en convaincre, nous faisons une analyse des groupes sur la base des caractéristiques des individus (cf. tableau 5.14). Comme nous pouvons le constater, il peut y avoir des hommes et des femmes issus des deux catégories socio-professionnelles au sein du même groupe, avec des taux techniques différents. En résumé, nous avons pu réaliser des agrégations plus intéressantes que celles que nous aurions pu obtenir uniquement sur la base des caractéristiques intrinsèques des assurés.

Proportion moyenne des femmes	Ecart type moyen de l'âge	Différentiel moyen du taux technique (max – min)
50%	25%	1.5%

Table 5.14. Caractérisation des groupes en rente d'éducation

Enfin, l'application de la méthode d'agrégation sur le portefeuille étudié a très peu d'impact sur les cashflows du passif projeté. En faisant la correspondance entre les caractéristiques des assurés et les groupes, nous pouvons modéliser notre portefeuille uniquement avec 79 Model Points.

	Provisions totales	dont PM	dont PSAP
Rente d'éducation	72 686	56 853	15 833
Erreur relative	**0.42%**	**0.57%**	**0.00%**

Table 5.15. Impact de l'agrégation sur les provisions en rente d'éducation

Rente de conjoint

Le décompte des possibilités en rente de conjoint est presque similaire à celui réalisé en rente de d'éducation, à deux détails près. La première différence réside dans l'intervalle de l'âge qui dans le cas de la rente de conjoint n'est pas limité, la deuxième est simplement le fait que la catégorie socio-professionnelle de l'assuré n'a pas d'incidence sur le calcul.

Le calcul du nombre total de cas (cf. tableau 5.16) montre que pour exploiter la base exhaustive, il faut appliquer les méthodes statistiques sur une base de données contenant 103 824 observations, caractérisées par les flux projetés jusqu'à l'horizon T. Nous faisons donc appel à une agrégation par le biais d'une base d'apprentissage plus restreinte. Pour ce faire, nous construisons d'abord des groupes sur la base du portefeuille réel. Ensuite, nous faisons une caractérisation des groupes identique à celle évoquée en rente d'éducation.

L'objectif de cette caractérisation des groupes est de simuler la base d'apprentissage en appliquant des perturbations sur le portefeuille réel, selon les règles suivantes :

Variable	Nombre de possibilités
Âge de 18 à 120 ans	(120-18+1) x 12 = 1236
Sexe	2
Taux technique	21 *(de 0% à 5% avec un pas de 0,25%)*
Fréquence de la rente	2 *(Mensuelle/Trimestrielle)*
Nombre total de possibilités	**103 824**

Table 5.16. Dimension de la base exhaustive en rente de conjoint

- L'âge des assurés étant une variable quantitative : nous générons alors des variables Gaussiennes selon les moyennes et les variances intra-groupe ;
- Pour les variables qualitatives (sexe, fréquence etc.) : nous générons une variable aléatoire multinomiale selon la distribution de la variable considérée au sein de chaque groupe.

Une base d'apprentissage contenant 70 000 individus est simulée selon cette méthode, puis les provisions mathématiques sont projetées grâce à notre modèle. Nous appliquons la méthode des K-means en faisant varier le nombre de groupes représentatifs jusqu'à ce que l'on obtienne une agrégation respectant le seuil d'erreur maximale.

Nombre de groupes	Erreur relative maximale	Erreur relative moyenne
250	0.26%	0.02%
150	0.78%	0.04%

Table 5.17. Application des K-means sur la base d'apprentissage en rente de conjoint

Compte-tenu du faible nombre de groupes créés, les scores des algorithmes d'apprentissage supervisé présentés dans le tableau ci-après sont plus élevés que ceux obtenus pour les autres garanties : en effet, le nombre d'observations moyen par groupe est plus élevé.

Méthode	Score moyen
Arbres de décision	98.95%
Random Forests	99.01%
Adaptive Boosting	99.23%

Table 5.18. Application des méthodes de classification supervisée - rente de conjoint

Finalement, l'agrégation du portefeuille réel est effectuée par le biais de l'algorithme

du Boosting adaptatif (conduisant au meilleur score lors de la validation croisée).

Nous passons de 3 916 assurés en portefeuille à 129 Model Points représentatifs avec un impact maîtrisé sur les provisions mathématiques.

	Provisions totales	dont PM	dont PSAP
Rente de conjoint	282 102	241 338	40 764
Erreur relative	**0.15%**	**0.17%**	**0.00%**

Table 5.19. Impact de l'agrégation sur les provisions en rente de conjoint

5.4 Récapitulatif de l'impact sur les provisions

L'application de la méthodologie proposée sur notre portefeuille a montré que le pouvoir d'agrégation des contrats dépend d'une part de la complexité de la modélisation et d'autre part du nombre de variables impactant la projection. En somme, **la dimension initiale de la base inventaire du passif est divisée par 9 avec une erreur relative très faible. Le temps de calcul est également réduit selon les mêmes proportions**.

	Provisions techniques	Nombres de lignes
Avant agrégation	2 220 689	66 212
Après agrégation	2 219 669	7 218
Erreur relative	0.03%	

Table 5.20. Impact de l'agrégation des Model Points

Finalement, nous avons vu que **l'efficacité de la méthode croît avec le nombre de paramètres entrant en jeu dans la modélisation de chaque garantie**. Pour la garantie rente éducation, qui requiert le plus grand nombre de paramètres parmi les risques étudiés, la méthode permet de diviser par 50 le nombre d'individus à modéliser. Par ailleurs, **l'analyse des regroupements réalisés fait ressortir une relative hétérogénéité au sein des groupes d'individus constitués** : un même individu représentatif peut représenter des bénéficiaires homme ou femme, avec des niveaux de taux techniques différents.

Chapitre 6

Impact de l'agrégation du passif sur le bilan économique

Nous disposons à présent des flux de passif issus de la projection du portefeuille réel ainsi que des flux liés à la projection du portefeuille agrégé. L'objectif de cette section est de présenter l'impact consolidé de l'agrégation du portefeuille sur le bilan économique Solvabilité II.

6.1 Évaluation du Best Estimate Liabilities

Pour comprendre l'impact de l'agrégation sur les engagements de l'assureur, nous utilisons la décomposition du suivante du BEL :

- Best Estimate de primes : il s'agit de la valeur actualisée des primes futures (au sens de la frontière des contrats Solvabilité II) diminuée de la valeur actualisée des prestations qui en découlent ;
- Best Estimate de frais : valeur actualisée des frais et des commissions relatifs à l'ensemble des contrats (i.e. stock et primes futures) ;
- Best Estimate de sinistres : cela correspond à la valeur actualisée des prestations futures probables associées aux contrats en portefeuille à la date d'évaluation ;
- Participation aux bénéfices : cet élément inclut le stock initial de provision d'égalisation et de réserve générale ainsi que la participation aux bénéfices future (Future Discretionary Benefits) ;
- Valeur terminale : elle inclut les provisions résiduelles et des plus ou moins-values latentes de fin de projection.

Notre agrégation du passif n'a pas d'impact sur le BE de primes dans le cadre de notre étude. Pour rappel, les lignes relatives aux primes futures sont modélisées par le biais de profils moyens. Cela signifie que le montant des primes projetées est multiplié

par le ratio sinistralité (S/P) pour obtenir la charge de sinistre, cette dernière est écoulée selon une hypothèse de cadences de liquidation.

L'hypothèse de S/P est renseignée à une maille fine (une hypothèse différente pour chaque garantie et pour chaque segment de clientèle). Le calcul conduit à un taux global implicite de 81.58% (incluant l'effet de l'actualisation des flux futurs aux taux Solvabilité II). Compte-tenu du montant des primes projetées (cf. 2.4), le montant de Best Estimate de primes est le suivant :

$$BE_{primes} = \text{- } 206\ 445$$

Un montant de Best Estimate de primes négatif signifie simplement que les primes payées par les assurés sont supérieures aux prestations. En revanche, le BE de prime n'inclut pas les frais engagés par l'entité.

L'impact de l'agrégation du passif en stock sur le BE de frais est une conséquence immédiate de l'impact sur les grandeurs sous-jacentes, à savoir les prestations et sur les provisions projetées. Dans le cadre de notre étude, nous adoptons les hypothèses suivantes de frais :

- 0.12% de frais de placements indexés sur la valeur de marché des actifs. Les placements au cours de la projection sont conditionnés par les provisions mathématiques projetées. De ce fait, l'impact sur les frais de placements est équivalent à celui obtenu pour les provisions mathématiques ;
- 4% de frais de gestion sur les prestations ;
- 8% de frais sur primes incluant les commissions et les frais d'administration, ils ne sont pas impactés par l'agrégation car le montant des primes reste inchangé.

	Avant agrégation	Après agrégation	% Erreur
BE_{frais}	216 532	216 433	0.05%

Table 6.1. Impact de l'agrégation sur le Best Estimate de frais

Analyse de l'impact sur le Best Estimate de sinistres

L'évaluation du Best Estimate de sinistres repose sur l'actualisation des prestations futures probables liées aux contrats en stock. Par rapport au calcul des provisions mathématiques, nous pouvons citer deux différences :

1– Contrairement au calcul des provisions mathématiques, l'actualisation ne se fait pas par le biais du taux technique mais plutôt sur la base des taux sans risque fournis par l'EIOPA ;

2– Pour la projection des prestations futures, le principe de prudence en Solvabilité I est remplacé par la meilleure estimation. Dans la pratique, cela signifie que les flux

de prestations peuvent être déterminés par le biais des tables d'expérience au lieu des tables réglementaires.

Dans le cadre de notre étude, seul le maintien de la garantie décès s'appuie sur deux tables différentes (une première table pour les provisions mathématiques prudentes et une deuxième pour les projections de type Best Estimate). De surcroît, la table d'expérience utilisée pour cette garantie est obtenue par abattement uniforme des tables de provisionnement pour tous les couples (âge,ancienneté). Cela signifie que les regroupements des assurés réalisés sur la base des tables réglementaires restent valables pour les projections Best Estimate également. De manière générale, si cette propriété n'est pas vérifiée, il suffit d'appliquer les méthodes de classification non-supervisée sur les prestations issues des tables d'expérience afin de construire les groupes représentatifs.

Au vu du critère de validation retenu, l'impact du changement des taux d'actualisation est maîtrisé. En effet, nous avons vu que la valeur absolue de l'erreur relative maximale sur les prestations est limitée à 1% sur toute la durée de la projection, ce qui signifie que l'erreur sur les flux actualisés est bornée par le même pourcentage, quel que soit le taux d'actualisation retenu.

En somme, l'impact sur le Best Estimate de sinistres (cf. tableau 6.2) est proche de l'impact observé sur les provisions Solvabilité I.

Garantie	BE de sinistres avant agrégation	BE de sinistres après agrégation	% Erreur
Invalidité	800 124	799 846	0.03%
Incapacité	447 193	448 631	0.32%
MGDC	111 836	111 176	0.59%
Rente d'éducation	76 070	75 672	0.52%
Rente de conjoint	294 430	293 690	0.25%
Santé	246 252	246 252	Profils moyens (non agrégés)
Temporaire décès	140 697	140 697	
Total	**2 116 602**	**2 115 965**	**0.03%**

Table 6.2. Impact de l'agrégation sur le Best Estimate de sinistres

Analyse de l'impact sur la participation aux bénéfices

Le montant initial de provision d'égalisation et de réserve générale au niveau de l'entité étudiée est de 62 756. Les dotations/reprises sont modélisées au cours de la projection selon les règles évoquées dans la présentation de l'approche de modélisation (cf. chapitre 3).

La projection des contrats est réalisée en run-off, cela signifie que les provisions augmentent en première année (projection d'un an de primes), et présentent une chronique

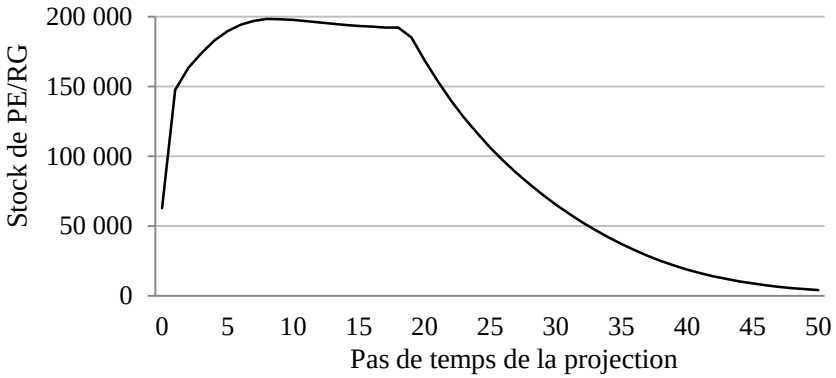
décroissante au cours de la projection. Cet aspect de modélisation a plusieurs conséquences sur la projection de la participation aux bénéfices.

Contractuellement, le montant de la provision d'égalisation (PE) est limité à 60% des cotisations prévoyance, le stock de PE est ainsi transféré en réserve générale au bout de la deuxième année de projection. De plus, nous limitons le stock de réserve générale au montant des provisions techniques projetées.

De ce fait, la projection du fonds de participation aux bénéfices peut se résumer en deux phases :

- ✓ Une phase de dotation en début de projection liée à la marge technique dégagée par l'année de primes projetées. Cette dotation s'explique également par la marge de prudence contenue dans les provisions des contrats en stock et finalement par la rémunération financière des provisions ;
- ✓ Une phase de transfert ou de sortie liée à la baisse des provisions techniques au cours de la projection. En pratique, ces transferts de PE/RG se matérialisent par une réduction des cotisations pour le client.

Le montant de participation aux bénéfices retenu dans le BEL correspond à l'actualisation aux taux sans risque des transferts de PE/RG réalisés au cours de la projection. Le montant de PE/RG résiduel est intégré dans la valeur terminale revenant au client.



Graphique 6.1. Évolution du stock de PE/RG au cours de la projection

La chronique du stock de PE/RG post-agrégation du passif n'est pas présentée dans le graphique précédent car elle est quasiment inchangée. De manière générale, l'agrégation du passif a un impact sur la participation aux bénéfices, principalement pour deux raisons :

1- La rémunération financière des provisions techniques peut varier car l'assiette de rémunération change. Dans notre cas, l'impact sur les provisions techniques est limité, ce qui conduit in fine à une très faible variation de la rémunération financière (inférieure à 1% tout au long de la projection) ;

2- Le calcul du solde technique avec un passif agrégé peut conduire à des écarts. Dans le cadre de notre étude, les soldes techniques sont supposés nuls sauf pour le maintien de la garantie décès (MGDC). Cela n'est pas problématique car la structure de la

table d'expérience est identique à celle de la table de provisionnement (pour rappel, la table d'expérience utilisée pour la garantie MGDC est obtenue par abattement uniforme de la table de provisionnement). Dans le cadre général, nous recommandons de prendre en compte les tables d'expérience dans la construction des groupes représentatifs pour maîtriser l'impact sur le solde technique des contrats en stock.

Pour finir cette section, l'impact de l'agrégation du passif sur la composante de participation aux bénéfices est présenté dans le tableau ci-après.

	Avant agrégation	Après agrégation	% Erreur
PB	261 152	261 066	0.03%

Table 6.3. Impact de l'agrégation sur la participation aux bénéfices

Synthèse des impacts sur le BEL

Nous avons vu que l'impact de l'agrégation du passif est très limité sur l'ensemble des composantes du Best Estimate Liabilities. C'est une conséquence directe du critère de validation retenu lors de l'application de la méthodologie d'agrégation. En effet, il s'agissait non seulement de contrôler l'impact sur la valeur actualisée des grandeurs projetées, mais aussi de limiter cet impact tout au long de la projection, car l'ensemble des flux interviennent dans la modélisation actif-passif ainsi que dans l'établissement des comptes de participation aux bénéfices. Le tableau ci-dessous fournit une synthèse des impacts sur le BEL.

	Avant agrégation	**Après agrégation**	**Valeur absolue de l'erreur relative**
BE de primes	-206 445	-206 445	0.00%
BE de sinistres	2 116 602	2 115 965	0.03%
BE de frais	216 532	216 433	0.05%
PB y compris FDB	261 152	261 066	0.03%
Valeur terminale	2 676	2 702	0.97%
BEL total	**2 390 517**	**2 389 721**	**0.03%**

Table 6.4. Synthèse des impacts consolidés sur le BEL

6.2 Évaluation du Solvency Capital Requirement

Nous nous appuyons sur la formule standard dont les modalités sont définies dans les actes délégués [C.E 2014] afin d'évaluer le capital de solvabilité requis (SCR).

Pour simplifier la présentation, nous omettons les modules de risques opérationnel et de contrepartie. D'une part, le SCR opérationnel se calcule via une formule fermée reposant sur les primes et les provisions Best Estimate : nous avons vu qu'il y a peu d'écart sur ces grandeurs à la suite de l'agrégation du passif. D'autre part, le calcul du risque de contrepartie repose sur l'exposition en réassurance et en créances, nous supposons que l'entité étudiée n'en dispose pas. À noter que la méthode d'agrégation proposée reste valable dans le cas général où certains contrats sont réassurés.

D'abord, nous nous appuyons sur la notion de Net Asset Value, cela correspond à la valeur de marché (VM) des actifs détenus par l'entité diminuée du BEL. De plus, nous utilisons la décomposition suivante de la NAV :

$$NAV = VM - BEL = VM - (BEG + PB) \tag{6.1}$$

Où le BEG correspond au Best Estimate Garanti, autrement dit, cette partie inclut tous les éléments du BEL évoqués précédemment hormis la participation aux bénéfices. Il convient de noter qu'il s'agit ici d'un abus de langage car la grandeur appelée communément BEG inclut également le stock de initial PB i.e. la partie non-discrétionnaire de la participation aux bénéfices.

Ensuite, pour un risque ψ, le montant du SCR_ψ relatif à ce risque sera obtenu par différence entre la NAV du scénario central et la NAV_ψ du scénario choqué : cela signifie que le modèle de projection est utilisé pour évaluer la variation des couples (VM_ψ,BEL_ψ) pour chaque scénario. Enfin, les matrices de corrélation fournies en formule standard nous permettront d'obtenir de montant final de SCR, tenant compte de la corrélation entre les risques.

Équation générale de l'impact SCR

Soient SCR_{init} et SCR_{fin} les montants des capitaux de solvabilité requis, pour un risque donné, avant et après agrégation. En utilisant la décomposition précédente, nous pouvons écrire :

$$\begin{cases} SCR_{init} = \Delta NAV_{init} = \Delta VM_{init} - (\Delta BEG_{init} + \Delta PB_{init}) \\ SCR_{fin} = \Delta NAV_{fin} = \Delta VM_{fin} - (\Delta BEG_{fin} + \Delta PB_{fin}) \end{cases} \tag{6.2}$$

Où ΔX représente la variation de la grandeur X entre le scénario central et le scénario

choqué et les indices *init* et fin représentent respectivement le passif avant agrégation et après agrégation.

Par ailleurs, l'agrégation du passif n'a aucun impact sur la valeur de marché des placements, en d'autres termes, nous avons $VM_{init} = VM_{fin}$ quel que soit le risque considéré. En effet, la valeur de marché varie uniquement en fonction du scénario économique étudié. En revanche, l'agrégation du passif a un impact sur la variation de la NAV par le biais du BEL et particulièrement par sa composante de participation aux bénéfices. De ce fait, l'impact de l'agrégation du passif sur le capital de solvabilité requis peut se résumer grâce à l'équation suivante :

$$SCR_{fin} - SCR_{init} = (\Delta BEG_{init} - \Delta BEG_{fin}) + (\Delta PB_{init} - \Delta PB_{fin}) \tag{6.3}$$

Dans la suite de cette section, nous réaliserons une analyse de l'impact de l'agrégation du passif pour les risques financiers et techniques sur la base de l'équation ci-dessus. Les modalités d'application des différents chocs sont détaillées en annexe.

Risques de marché

Pour commencer nous traitons d'abord les chocs financiers n'ayant pas d'impact sur le taux d'actualisation du passif, à savoir les risques actions, immobilier et le spread de crédit. Pour ces chocs, l'équation 6.3 devient :

$$SCR_{fin} - SCR_{init} = \Delta PB_{init} - \Delta PB_{fin} \tag{6.4}$$

Car le BEG n'est pas impacté par la variation de la valeur de marché des actifs concernés par les chocs cités précédemment ($\Delta BEG = 0$ avant et après agrégation). Seule la participation aux bénéfices change par le biais de la variation du taux de rendement actuariel. Pour rappel, la rémunération des provisions techniques dans les comptes de PE/RG se fait selon le TRA de l'entité.

L'impact observé sur le TRA après l'agrégation du passif est très faible (un impact moyen de 0.02% sur l'horizon de projection) car ce taux subit uniquement un effet de second ordre. En effet, la chronique des rendements actuariels change très légèrement parce que les provisions techniques en face des placements sont impactées suite à l'agrégation, mais nous avons vu précédemment que cela est très marginal.

En outre, pour les chocs de taux à la baisse et à la hausse, il existe une deuxième source d'écart sur le SCR due à l'actualisation des flux futurs selon les taux des scénarios choqués. Cependant, au vu du critère de validation fixé, l'écart relatif sur les prestations futures est inférieur à 1% quel que soit le pas de temps de la projection. Cela est vrai aussi dans les scénarios de choc car les flux de prestations en prévoyance et santé sont insensibles à la variation des conditions de marché.

En résumé, l'impact final sur le SCR de marché est très faible, les ordres de grandeur sont indiqués dans le tableau ci-après.

	Avant agrégation	Après agrégation	% Erreur
SCR marché	347 709	347 725	0.005%

Table 6.5. Impact de l'agrégation sur le SCR marché

Risques de souscription vie et santé

En examinant les modalités d'application des chocs de souscription en formule standard sur le périmètre étudié (prévoyance et santé), nous pouvons distinguer deux grandes catégories :

Catégorie 1 : les SCR obtenus par formule fermée, c'est le cas par exemple des chocs primes et réserves dont le calcul se fait sur la base des primes et du Best Estimate de sinistres. Cette première catégorie n'impacte pas les comptes de participation aux bénéfices et peut être obtenue sans réévaluation des cashflows futurs. De ce fait, **l'impact de l'agrégation du passif sur cette première catégorie est maîtrisé car il dépend simplement de l'écart observé sur les grandeurs sous-jacentes, à savoir le BEL ou les provisions techniques** ;

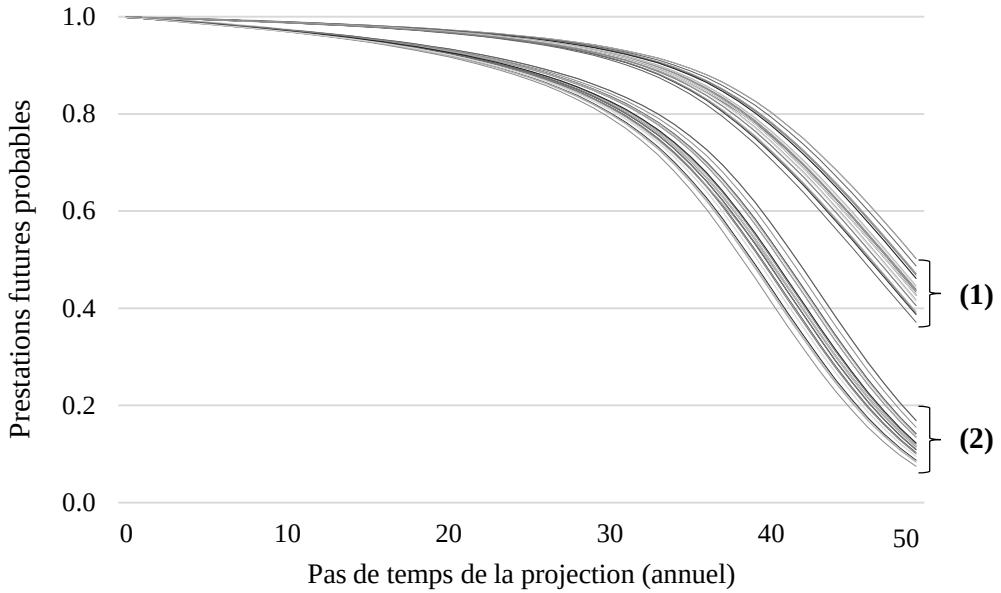
Catégorie 2 : les SCR obtenus par le biais de chocs additifs ou multiplicatifs appliqués sur les tables de mortalité, celles de passage de l'état d'incapacité à l'état d'invalidité ainsi que celles de maintien en incapacité/invalidité. L'application de ces chocs conduit à un impact sur le solde technique, faisant varier le montant de la participation aux bénéfices. En conséquence, cette catégorie de chocs nécessite de réaliser la projection Actif-Passif dans le cadre du scénario choqué.

En ce qui concerne la deuxième catégorie, force est de constater que ces chocs ne changent pas fondamentalement la structure des tables utilisées dans le scénario central. En d'autres termes, l'application de la méthode d'agrégation aux cashflows de passif choqués conduit à la construction de Model Points identiques à ceux obtenus en scénario central. Cela est vérifié car les tables choquées correspondent simplement à une transformation affine des tables utilisées dans le scénario central.

Par exemple, le risque de longévité concerne les rentiers en portefeuille, il consiste à choquer à la baisse la table de mortalité selon un facteur multiplicatif de 80%, uniforme pour tous les âges. De ce fait, si les chroniques de prestations d'un groupe d'assurés sont proches dans le scénario central, alors elles le sont également en scénario choqué, car elles sont simplement obtenues selon un facteur multiplicatif.

Pour illustrer cet aspect, on considère l'exemple d'un groupe d'assurés en rente de

conjoint constitué de 26 individus. Pour rappel, nous avons construit nos groupes en appliquant les K-means sur les flux du passif issus de la projection en scénario central. Dans l'objectif d'obtenir ces flux dans le scénario de longévité, nous appliquons un choc multiplicatif à la baisse sur les tables de mortalité comme le montre le graphique suivant :



Graphique 6.2. Comparaison des prestations d'un groupe d'assurés bénéficiant d'une rente de conjoint : **(1)** dans le scénario de longévité - **(2)** dans le scénario central

Comme nous pouvons le voir, les chroniques des prestations des différents assurés appartenant au groupe considéré restent proches après l'application du choc de longévité. Cela signifie que le groupe créé sur la base du scénario central reste valable également dans le scénario choqué.

Pour conclure, l'impact de la méthode d'agrégation sur le SCR de souscription est négligeable, le niveau de l'erreur obtenue en calculant ce module de SCR avant et après agrégation confirme les résultats de l'analyse menée précédemment.

	Avant agrégation	Après agrégation	% Erreur
Vie	10 068	10 087	0.19%
Santé	291 147	291 742	0.20%

Table 6.6. Impact de l'agrégation sur les SCR de souscription vie et santé

SCR total

Nous avons vu que l'impact sur les différents modules de SCR est faible. Il convient à ce stade de calculer le SCR consolidé grâce aux coefficients de corrélation entre les risques prévus par la formule standard. Les trois modules de marché, de souscription vie et santé sont corrélés à hauteur de 25%. Posons $\Gamma = \{\text{Marché, Santé, Vie}\}$, nous avons la formule suivante :

$$SCR = \sqrt{\sum_{i\in\Gamma} SCR_i^2 + \sum_{(i,j)\in\Gamma} 25\% \times SCR_i \times SCR_j} \tag{6.5}$$

Sur la base de cette formule, nous avons les montants suivants de SCR :

	Avant agrégation	Après agrégation	% Erreur
SCR	509 512	509 976	0.09%

Table 6.7. Impact de l'agrégation sur le SCR total

6.3 Impact consolidé sur le bilan Solvabilité II

Évaluation de la Risk Margin

En Solvabilité II, les provisions techniques sont évaluées en vision économique, cela signifie qu'elles correspondent à la valeur de transfert du portefeuille vers un autre assureur. De ce fait, le BEL est ajusté d'une marge pour risque ou Risk Margin définie dans [C.E 2014]. Nous utilisons la méthode du coût du capital afin d'évaluer la marge pour risque (RM) :

$$RM = Duration \times CoC \times SCR \tag{6.6}$$

D'une part, la duration de Macaulay [1] calculée sur la base des flux du BEL est de 8 ans, elle est stable après l'agrégation du passif. D'autre part, le taux du coût du capital (CoC) retenu est de 6%. Cela nous conduit donc aux montants présentés dans le tableau 6.8.

1. Pour le vecteur de cashflows actualisés $(\widetilde{CF}_t)_{0\leq t\leq T}$, la duration de Macaulay est donnée par $\frac{\sum_{t=1}^{T} t\times\widetilde{CF}_t}{\sum_{t=1}^{T}\widetilde{CF}_t}$

	Avant agrégation	Après agrégation	% Erreur
RM	244 566	244 789	0.09%

Table 6.8. Impact de l'agrégation sur la Risk Margin

Bilan économique

En rassemblant les différentes composantes nécessaires, notamment la valeur de marché des placements, le BEL et la Risk Margin, nous pouvons mesurer les fonds propres économiques et représenter le bilan de l'entité étudiée avant et après agrégation des contrats en portefeuille.

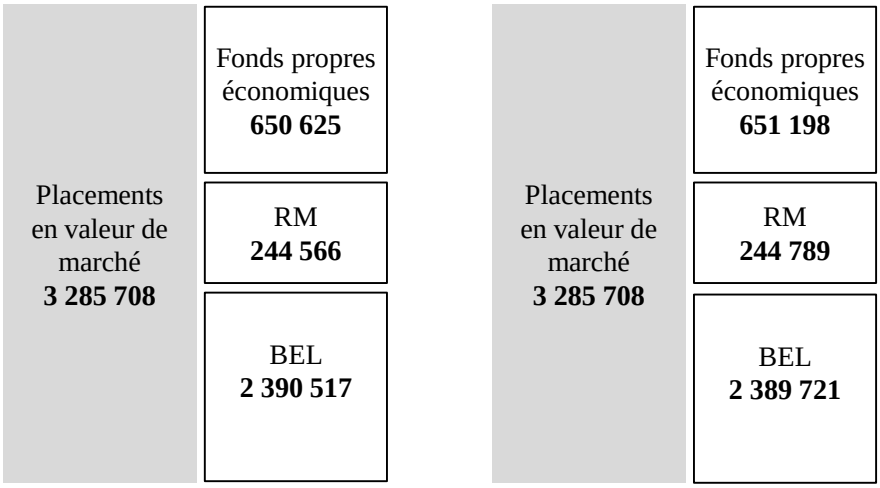


Figure 6.1. Bilan Solvabilité II final

Comme nous pouvons le constater, l'impact sur les fonds propres économiques est de l'ordre de 0.09%. Le ratio de solvabilité peut être obtenu en rapportant les fonds propres économiques au capital de solvabilité requis SCR. Cela nous conduit quasiment au même ratio avant et après agrégation :

	Avant agrégation	Après agrégation
Ratio de solvabilité	127.696%	127.692%

Table 6.9. Impact de l'agrégation sur le ratio de couverture Solvabilité II

Pour rappel, la taille du fichier d'inventaire du passif projeté est divisée par 9. Le gain en temps de calcul est davantage important lors du calcul du SCR car certains chocs nécessitent de réévaluer les cashflows dans le cadre de chaque scénario choqué.

Chapitre 7

Limites, améliorations possibles et perspectives

L'objectif de cette partie est d'identifier les limites de la méthode d'agrégation proposée ainsi que les éléments pouvant rendre son utilisation possible dans un cadre plus général. Plus précisément, nous discuterons deux grands aspects :

1– **Hypothèses :** l'objectif est de challenger les hypothèses retenues lors de la mise en place de la méthode d'agrégation et particulièrement les hypothèses pouvant restreindre la généralisation de la méthode ;

2– **Méthodologie :** il s'agira de creuser une composante primordiale retenue dans méthode d'agrégation qui est la distance Euclidienne comme mesure de similarité entre les assurés.

Hypothèses : modélisation de la participation aux bénéfices

Dans le cadre de notre étude, les comptes de PB sont projetés via une configuration moyenne pour l'ensemble des contrats en portefeuille. Ce choix de modélisation a été retenu car les clauses de participations aux bénéfices ne présentent pas une grande variabilité au sein de notre portefeuille, nous nous sommes appuyés donc sur un paramétrage unique de la PB pour la modélisation prospective de l'ensemble des contrats, notamment un indice financier de référence unique et un taux de partage moyen pour la rémunération des provisions.

Dans le cas général, cette hypothèse de faible variabilité des clauses de PB entre les différents contrats peut ne pas être vérifiée lorsque l'assureur considéré détient une part importante de contrats participatifs. Ceci dit, à défaut de pouvoir modéliser les comptes

de participation aux bénéfices par une configuration unique, il suffit d'appliquer notre méthode d'agrégation pour chaque configuration présente dans le portefeuille.

Plus précisément, il suffira d'appliquer l'algorithme de prédiction des Model Points représentatifs sur des sous-groupes définis par une clause différente de participation aux bénéfices. Cela réduit de fait le pouvoir d'agrégation de la méthode mais la taille du portefeuille de contrats après agrégation restera limitée pour peu que le nombre de clauses de PB ayant des caractéristiques différentes soit faible. Ce dernier point est davantage vérifié lorsqu'il s'agit d'un portefeuille contenant des contrats sans clauses de PB ou des contrats collectifs contenant un nombre important d'assurés.

Hypothèses : structures des tables d'expérience et de provisionnement

La deuxième spécificité des hypothèses retenues dans le cadre de nos travaux concerne les tables de mortalité et les lois adoptées en incapacité et en invalidité. En effet, pour la majorité des garanties, les tables d'expérience et de provisionnement sont identiques, autrement dit, les provisions Solvabilité I sont calculées selon les mêmes lois que celles utilisées pour le BEL. Cela n'est pas toujours vérifié car l'objectif des projections Solvabilité II est de calculer la meilleure estimation des provisions et non des provisions prudentes. De surcroît, la table d'expérience utilisée pour le maintien de la garantie décès avait exactement la même structure que la table de provisionnement.

Cependant l'utilisation d'une table d'expérience ayant la même structure que la table de provisionnement n'a pas d'impact sur notre méthodologie d'agrégation, comme vu dans la section 6.2. Ainsi, dans le cas général où la table d'expérience n'a pas la même structure que la table de provisionnement, nous recommandons d'utiliser les cashflows de passif issus des deux tables lors de la construction des groupes représentatifs : cela permet de maîtriser l'impact de l'agrégation sur la marge de prudence reconnue en fonds propres dans le bilan Solvabilité II.

Il est en revanche important de noter que l'utilisation des flux de prestations issus des deux projections (provisions et expérience) conduit à multiplier par deux le nombre de variables utilisées lors de la construction des groupes. Nous disposerons donc de $2 \times T$ variables caractérisant chaque assuré de la base d'apprentissage, où T représente l'horizon de la projection.

Pour limiter le nombre de variables, une amélioration possible consiste à appliquer une Analyse en Composantes Principales (ACP) sur les données avant de construire les groupes représentatifs. Cela permet en effet de détecter les dépendances linéaires qui peuvent exister entre les cashflows de passif et donc de réduire le nombre de variables.

Généralement, les probabilités de décès d'une année à l'autre ne présentent pas de corrélation linéaire. Les probabilités de décès sont plutôt obtenues selon des modèles exponentiels. De ce fait, l'application du logarithme aux cashflows de prestations avant d'inférer les dépendances linéaires peut s'avérer dans certains cas très efficace.

Pour s'en convaincre, la figure 7.1 montre clairement que l'application du logarithme sur les flux de prestations utilisés pour la construction des groupes en maintien de la garantie décès améliore considérablement la variance expliquée en fonction du nombre de variables retenues.

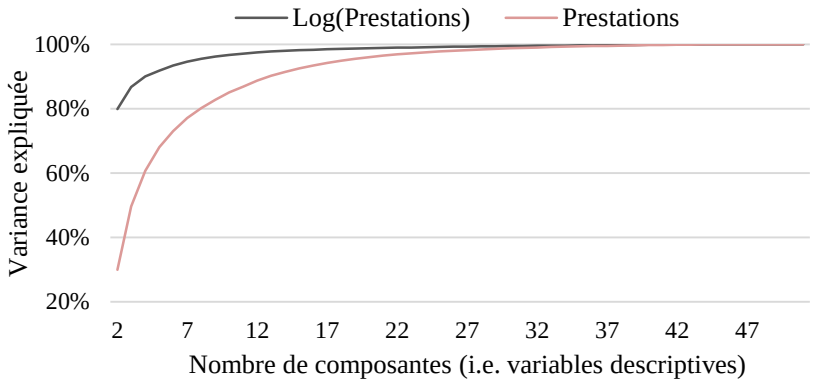


Figure 7.1. Réduction de la dimension de la base d'apprentissage grâce à l'analyse en composantes principales

Méthodologie : mesure de distance entre les individus

Notre méthode de classification non-supervisée s'applique sur les flux futurs probables : chaque individu i dans notre base de données est caractérisé par une chronique de flux $CF_t(i)$ où t varie entre 0 et l'horizon de projection T. En d'autres termes, les assurés sont caractérisés par des chroniques de prestations ou de provisions pouvant être interprétées comme des séries temporelles.

L'article de recherche [Najmeddine 2012] fournit des éléments très intéressants concernant les mesures de similarité pouvant être utilisées dans le cadre de l'agrégation des séries temporelles. En effet, c'est une problématique assez classique qui revient dans plusieurs domaines. Dans l'article cité, il s'agit d'agréger des séries temporelles issues des données énergétiques de bâtiments.

La mesure de similarité *The Complexity Invariance Distance (CID)* peut s'avérer très efficace lorsqu'il s'agit d'agréger des séries temporelles présentant un caractère irrégulier. En effet, dans le cadre de notre étude, les prestations de la garantie santé sont modélisées selon des profils moyens et ne sont donc pas concernées par l'agrégation. Dans un cadre

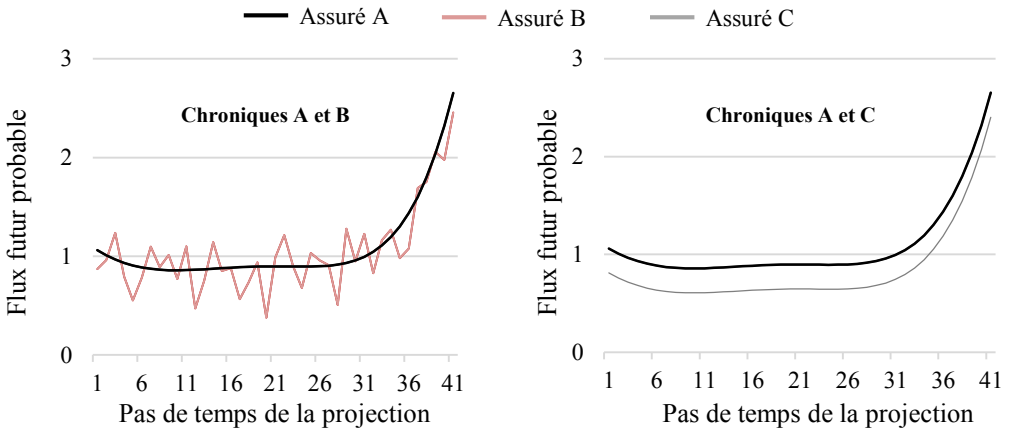
plus général, les prestations en santé pour un assuré peuvent présenter une chronique très irrégulière.

Soit x_{i_1} et x_{i_2} les casfhlows de passif caractérisant deux assurés i_1 et i_2, la distance CID est donnée par la formule suivante :

$$CID(i_1, i_2) = ||CF(i_1) - CF(i_2)||_2 \times \frac{max(CE(i_1), CE(i_2))}{min(CE(i_1), CE(i_2))} \tag{7.1}$$

Où $CE(i) = \sqrt{\sum_{t=2}^{T}(CF_t(i) - CF_{t-1}(i))^2}$ l'estimation de la complexité de la série temporelle des flux de l'assuré i.

Comme le montre la figure ci-après et contrairement à la distance Euclidienne, l'avantage de cette mesure de similarité est de pouvoir identifier les irrégularités qui peuvent exister dans une chronique de prestations.



Distance Euclidienne entre A et B	1.6
Distance CID entre A et B	2.13

Distance Euclidienne entre A et C	1.79
Distance CID entre A et C	1.79

Figure 7.2. Intérêt de l'utilisation de la mesure de similarité CID

L'objectif est de construire deux groupes à partir des trois assurés A, B et C. L'utilisation de la distance Euclidienne conduit à agréger au sein du même groupe les deux individus du graphe à gauche, et ce malgré la forte irrégularité de la chronique des prestations de l'assuré B. En revanche, l'utilisation de la distance CID conduit à l'agrégation suivante des trois individus en deux groupes :

$$Groupe_1 = \{A, C\} \text{ et } Groupe_2 = \{B\}$$

Conclusion

Contrairement à ce que l'on pourrait penser, regrouper les assurés ayant des caractéristiques (âge, sexe etc.) similaires n'est pas toujours une approche optimale lorsqu'il s'agit de traiter la problématique de l'agrégation du passif. Il existe en effet des critères de similarité permettant de réaliser des agrégations plus intéressantes et de réduire significativement la taille du passif à projeter, notamment en s'appuyant sur les flux futurs probables des assurés.

A la suite de notre étude, et sur la base des résultats obtenus sur le portefeuille de prévoyance et santé étudié, la méthodologie proposée permet de réduire considérablement le nombre de contrats à projeter sans perte significative de précision. De plus, en utilisant les algorithmes d'apprentissage automatisé, il a été possible de rendre notre méthode stable dans le temps, c'est-à-dire insensible au vieillissement et à l'évolution du passif d'une année à l'autre.

Si le critère de validation proposé permet de limiter l'erreur de l'agrégation du passif à un seuil fixé en amont, il n'en demeure pas moins que l'application de la méthode dans le cadre de la production réglementaire requiert une maîtrise approfondie des impacts et un back-testing régulier : en effet, l'autorité de contrôle peut imposer une marge de solvabilité complémentaire *(Capital add-on)* si le regroupement des assurés introduit un biais dans les résultats.

L'agrégation du passif offre beaucoup d'avantages d'un point de vue opérationnel. En outre, cela permet de réduire significativement les coûts des plateformes informatiques nécessaires à la modélisation et la projection des contrats. C'est d'ailleurs un sujet d'actualité dans un contexte réglementaire plus complexe, notamment avec l'entrée en vigueur de la norme IFRS17.

Annexes

Annexe I – Complexité des méthodes de classification non-supervisée

Lors de la présentation des méthodes de classification non-supervisée ou de *Clustering*, nous avons comparé la performance de quatre méthodes classiques sur des bases de données synthétiques ayant les mêmes caractéristiques que les bases réelles. Nous avons vu que la méthode des K-means a la meilleure performance. Cela est bien en ligne avec la complexité temporelle de chaque méthode, comme le montre le tableau ci-dessous.

Méthode	**Complexité temporelle**
K-means	$O(NK)$
Classification ascendante hiérarchique	$O(N^3)$
DBSCAN	$O(N^2)$
Modèle de mélange Gaussien	$O(NK^2)$

Table 7.1. Complexité des méthodes de Clustering

Où N représente le nombre d'observations et K le nombre de classes à former.
À noter que d'autres paramètres impactent la complexité temporelle de ces méthodes (notamment le nombre d'itérations nécessaires pour la convergence des algorithmes) : seuls les impacts des paramètres N et K sont considérés dans cette présentation.

Annexe II – Algorithme du Boosting adaptatif

En classification supervisée, nous nous sommes intéressés à l'algorithme SAMME *(Stagewise Additive Modeling using a Multi-class Exponential loss function)* introduit par [Zhu 2006].

Il s'agit en effet d'une généralisation de la méthode AdaBoost (applicable uniquement en classification binaire) à la problématique de l'apprentissage automatisé à classes multiples. Nous reprenons les notations de la section 4.2.6 pour présenter l'algorithme :

Algorithme 1 SAMME

Initialisation: Affectation des poids $\omega_i = \frac{1}{N}$ à l'ensemble des observations $i \in [1, N]$

Pour $m = 1...M$

(1) Construction de l'arbre $\mathcal{T}_m$ sur la base des données d'apprentissage pondérées selon les poids ω_i

(2) Calcul de la proportion des observations mal classées

$$\epsilon(m) = \frac{\sum_{i=1}^{N} \omega_i . \mathbb{1}(C_i \neq \mathcal{T}_m(x_i))}{\sum_{i=1}^{N} \omega_i}$$

(3) Calcul du coefficient de pondération

$$\alpha_m = \log\left(\frac{1-\epsilon(m)}{\epsilon(m)}\right) + \log(M-1)$$

(4) Calcul des nouveaux poids, pour $i \in [1, N]$

$$\omega_i \leftarrow \omega_i . \exp\left(\alpha_m . \mathbb{1}(C_i \neq \mathcal{T}_m(x_i))\right)$$

(5) Re-normalisation des ω_i

Sortie: La classe C_x d'un nouvel individu x est donnée par :

$$C_x = \underset{C}{\operatorname{argmax}} \sum_{m=1}^{M} \alpha_m . \mathbb{1}(\mathcal{T}_m(x) = C)$$

Annexe III – Modalités d'application des chocs SCR

Le tableau 7.2 fournit les détails concernant les chocs de la formule standard appliqués au portefeuille à disposition. Il s'agit d'un complément à la section 6.2 présentant les montants des SCR ainsi que leur variation suite à l'agrégation du passif. Les modalités d'application des chocs sont disponibles dans le règlement délégué de la Commission Européenne [C.E 2014] et plus précisément dans les articles cités dans le tableau ci-après.

Module	Risque		Référence actes délégués de l'UE
SCR Marché	Sensibilité du portefeuille au niveau / à la volatilité de la valeur de marché des **actions**		Art. 168-173
	Risque lié aux changements affectant la courbe des **taux** d'intérêt (à la baisse ou à la hausse)		Art. 165-167
	Sensibilité au niveau des marges (**spread**) de crédit par rapport à la courbe des taux d'intérêt sans risque		Art. 175-181
	Risque résultant de la volatilité des prix de marché de l'**immobilier**		Art. 174
SCR Souscription Vie	Choc de **mortalité** permanent pour le maintien de la garantie décès ainsi que pour les temporaires décès		Art. 137
	Augmentation permanente de la **longévité** des rentiers (rente d'éducation et rente de conjoint)		Art. 138
	Hausse des **frais** et de l'inflation pour l'ensemble des garanties « vie » sur l'horizon de la projection		Art. 140
	Choc additif sur les taux de décès appliqué uniquement en première année de projection (**catastrophe vie**)		Art. 143
SCR Souscription Santé	SLT*	Augmentation de la **longévité** en invalidité tout au long de la projection	Art. 153
		Hausse des **frais** et de l'inflation pour l'ensemble des garanties « non vie » sur l'horizon de la projection	Art. 157
		Augmentation des montants des rentes d'invalidité (choc de **révision**)	Art. 158
	Non SLT	Risque de **primes** (volatilité affectant les sinistres) et **réserves** (volatilité affectant le règlement des sinistres)	Art. 146-149
	CAT	Choc **catastrophe** composé des trois sous-risques: accident de masse, concentration d'accidents, pandémie	Art. 160-163

Table 7.2. Modules de SCR impactant le portefeuille étudié

(*) Garanties similaires à la vie, sur le portefeuille considéré, cela correspond à la garantie invalidité. Les garanties Non SLT correspondent quant à elles à l'incapacité et au remboursement des frais de soin.

Il convient de mentionner que la liste des risques cités dans le tableau n'est pas exhaustive : certains chocs sont omis car ils ne présentent pas d'intérêt dans le cadre de notre étude portant sur l'agrégation du passif.

Bibliographie

[Arthur 2007] D. Arthur et S. Vassilvitskii. *K-means++*, 2007.

[Breiman 1984] L. Breiman, J. Friedman, R. Olshen et C. Stone. *Classification and Regression Trees*, 1984.

[Breiman 2001] L. Breiman. *Random forests, Machine Learning*, 2001.

[C.E 2014] Commission Européenne C.E. *Règlement Délégué Article 35*, 2014. Journal officiel de l'Union européenne.

[Chawla 2002] N.V. Chawla, K.W. Bowyer, L.O. Hall et W.P Kegelmeyer. *Synthetic Minority Over-sampling Technique*, 2002.

[Cohen 1960] J. Cohen. *A coefficient of agreement for nominal scales*, 1960.

[Couloumy 2013] A. Couloumy. *Critères d'agrégations de jeu de données passif pour les calculs sous la Formule Standard de Solvabilité 2*, 2013. Mémoire d'actuariat.

[EIOPA 2014] EIOPA. *EIOPA-BoS-14/166 FR*, 2014. Orientations sur la valorisation des provisions techniques.

[Ester 1996] M. Ester, H.P. Kriegel, J. Sander et X. Xu. *A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise*, 1996.

[Freund 1996] Y. Freund et R.E. Schapire. *New Boosting Algorithm*, 1996.

[Goffard 2015] Pierre-Olivier Goffard. *Is it optimal to group policyholders by age, gender, and seniority for BEL computations based on model points ?*, 2015. Thèse.

[Le Corre 2012] Erwan Le Corre. *Constitution de groupes homogènes de risque dans le cadre de Solvabilité II*, 2012. Mémoire d'actuariat.

[MacQueen 1967] J. MacQueen. *Some Methods for Classification and Analysis of Multivariate Observations*, 1967.

[Najmeddine 2012] H. Najmeddine, F. Suard, A. Jay, P. Marechal et M. Sylvain. *Mesures de similarité pour l'aide à l'analyse des données énergétiques de bâtiments*, 2012.

[Tann 2011] Philippe Tann. *Optimisation du niveau de segmentation des portefeuilles de prévoyance individuelle sous Solvabilité II*, 2011. Mémoire d'actuariat.

[Zhu 2006] J. Zhu, S. Rosset, H. Zou et T. Hastie. *Multi-class AdaBoost*, 2006.

Note de synthèse

Mise en contexte

Le calcul des provisions Best Estimate nécessite de projeter les flux futurs probables relatifs à l'ensemble des assurés et des bénéficiaires du portefeuille. L'approche de projection en tête par tête s'avère très contraignante, compte-tenu des temps de calcul importants ainsi que des ressources informatiques considérables nécessaires à l'utilisation des modèles de projection.

L'article 35 du règlement délégué (UE) 2015/35 de la Commission Européenne dispose que pour le calcul de la meilleure estimation des provisions, les assureurs peuvent recourir à des méthodes d'agrégation des passifs, à condition que ces derniers présentent une nature de risque identique et que l'agrégation n'ait pas d'impact significatif sur les résultats.

L'agrégation des passifs présente par ailleurs l'intérêt de réaliser des exercices ORSA ainsi que des études ALM de manière beaucoup plus fluide, avec un passif léger et facilement exploitable. En permettant de réduire les coûts liés aux plateformes informatiques de modélisation actuarielle, elle constitue un outil de performance opérationnelle auquel la plupart des assureurs recourent actuellement.

Le présent mémoire propose une approche d'agrégation appliquée à un portefeuille de prévoyance et santé. Il convient de noter que les principes de l'approche d'agrégation proposée pourraient s'appliquer en toute généralité à des portefeuilles et risques différents (épargne, emprunteur, etc.).

La bibliographie relative à la problématique de l'agrégation des passifs est riche et concerne différentes typologies de portefeuilles. En prévoyance et santé, les méthodes existantes reposent sur les caractéristiques des assurés et des bénéficiaires (par exemple l'âge et le sexe). Il s'agit de regrouper des individus dont les caractéristiques sont proches : par exemple, deux femmes nées la même année et bénéficiant chacune d'une rente de conjoint, sont modélisées avec un seul individu représentatif.

La méthode proposée dans ce mémoire s'appuie sur une approche différente : plutôt que de chercher à regrouper des individus en fonction de leurs caractéristiques, il s'agit de créer des groupes en fonction des cashflows probables qui leur sont associés. Cette différence d'approche nous permettra de réaliser des regroupements d'assurés plus intéressants que ceux obtenus par les méthodes reposant sur les caractéristiques des individus. Pour s'en convaincre, prenons le cas d'une garantie temporaire décès à horizon 1 an modélisée avec la table TH/TF00-02 et les deux individus ci-dessous :

Assuré	Probabilité de décès dans l'année
Femme d'âge 46 ans	$0,1951\%$
Homme d'âge 38 ans	$0,1953\%$

Tableau 7.1. Illustration du critère de similarité entre deux assurés

Ces deux individus génèrent des flux futurs probables similaires car les probabilités de décès sont très proches : il est donc possible de les regrouper sans perte significative de précision, ce qui est typiquement réalisé avec la méthode présentée dans le mémoire. Un tel regroupement ne serait clairement pas possible avec les méthodes reposant sur les caractéristiques intrinsèques des individus compte-tenu de la différence d'âge et de sexe.

Présentation du portefeuille étudié

Le tableau ci-dessous présente les garanties du portefeuille étudié. Il s'agit d'un portefeuille classique d'une entité commercialisant des contrats collectifs de prévoyance et santé.

Garantie	Nature de l'engagement
Incapacité temporaire	Indemnités journalières d'arrêt de travail en cas d'incapacité temporaire
Invalidité permanente	Versement d'une rente mensuelle en cas d'invalidité permanente
Santé	Remboursement des frais de soins de santé
Temporaire décès	Temporaire décès (versement d'un capital) – contrats annuels
Maintien de la garantie décès	Maintien des garanties décès suite à un arrêt de travail
Rente d'éducation	Versement d'une rente d'éducation à l'enfant en cas de décès du parent assuré
Rente de conjoint	Versement d'une rente au conjoint bénéficiaire en cas de décès de l'assuré

Tableau 7.2. Risques et garanties offertes aux assurés au sein du portefeuille étudié

La méthode d'agrégation proposée porte sur les passifs relatifs aux engagements liés à des sinistres connus à la date d'inventaire. Elle s'applique à toutes les garanties listées précédemment à l'exception des garanties temporaire décès et santé, car ces dernières sont modélisées selon des profils moyens.
Les engagements portent soit sur le versement de rentes à des individus bénéficiaires, soit sur le versement d'un capital décès en cas de décès de l'individu assuré. Les flux futurs probables associés à ces individus sont obtenus en utilisant :

- D'une part, des hypothèses techniques, notamment des probabilités de décès ou de maintien dans l'état d'incapacité/invalidité ;
- D'autre part, un outil de modélisation actuarielle. L'approche de modélisation retenue dans le cadre de notre étude repose sur un modèle de changement d'états,

autrement dit, les flux futurs probables sont obtenus en réalisant la projection des effectifs dans chaque état (validité, incapacité, invalidité et décès).

Les caractéristiques des individus qui doivent être renseignées pour pouvoir projeter les flux futurs probables, ainsi que les hypothèses techniques sont détaillées ci-dessous :

Garantie	Paramètres impactant la projection	Hypothèses sous-jacentes
Incapacité / Invalidité	- Âge de l'assuré - Date d'entrée dans l'état d'incapacité ou d'invalidité	- Tables du BCAC de maintien en incapacité/invalidité et de décès, sans distinction en fonction du sexe
Maintien de la garantie décès		
Rente d'éducation	- Âge et sexe du bénéficiaire - Catégorie socio-professionnelle du parent - Taux technique - Fréquence de la rente	- Table de mortalité TGH/TGF-05 - Loi d'expérience de poursuite des études par catégorie socio-professionnelle
Rente de conjoint	- Âge et sexe du bénéficiaire - Taux technique - Fréquence de la rente	- Table de mortalité TGH/TGF-05

Tableau 7.3. Paramètres et hypothèses de modélisation par garantie

L'objectif de la méthode d'agrégation proposée est de réduire le nombre d'individus qui sera finalement modélisé par rapport au nombre d'individus en portefeuille, en maîtrisant l'erreur induite sur les résultats. Dans la suite, les individus présents dans la base agrégée seront désignés par le terme « d'individus représentatifs ». Par exemple, un individu représentatif qui représenterait trois rentiers de la base d'origine, aurait des caractéristiques propres et un montant de rente égal à la somme des montants de rente des trois rentiers.

Construction d'individus représentatifs

Dans cette partie, la base de données exploitée est constituée d'une liste d'individus, avec pour chacun d'entre eux ses caractéristiques et les cashflows futurs associés.

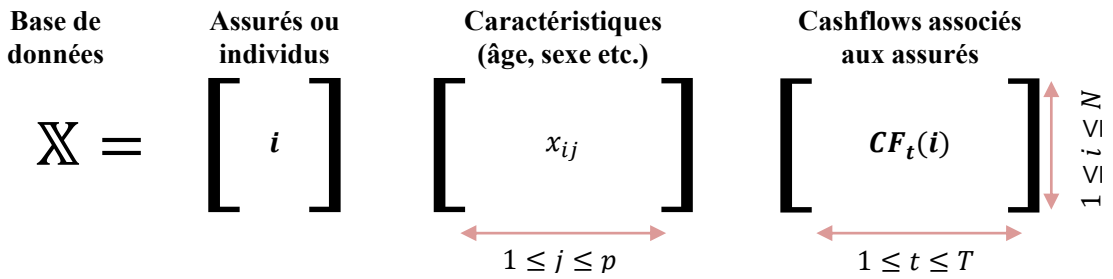


Schéma 7.1. Base de données considérées

Concrètement, les cashflows futurs correspondent soit aux prestations futures probables soit aux provisions mathématiques projetées, en fonction de la garantie modélisée.

La première étape de notre méthode d'agrégation consiste à construire des individus candidats à représenter les individus présents dans la base de données considérée.

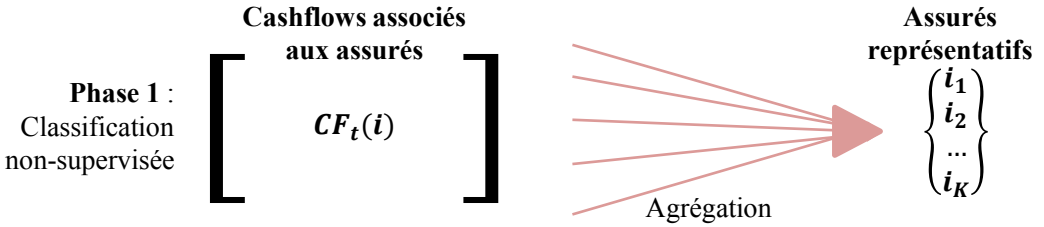


Schéma 7.2. Exploitation des méthodes de classification non-supervisée

La construction des individus représentatifs est réalisée avec une méthode de classification non-supervisée, la méthode des K-means. L'application de cette méthode nécessite de définir plusieurs paramètres qui sont présentés en détail dans le mémoire. Seul le critère de validation de l'agrégation est présenté ci-dessous.

A chaque pas de projection t, les cashflows du passif ont une incidence sur les comptes de participation aux bénéfices ainsi que sur l'interaction Actif-Passif. Dès lors, le critère de validation retenu a été construit afin de limiter le niveau d'erreur à chaque pas de temps. Notons, pour un pas de temps t, CF_t la somme des cashflows d'un groupe d'individus avant agrégation et CF_t^{ag} les cashflows obtenus après agrégation du groupe. Le critère de validation de l'agrégation est le suivant :

$$\max_{t\in[1,T]} \left| \frac{CF_t - CF_t^{ag}}{CF_t} \right| < 1\% \tag{7.2}$$

Avec ce critère, nous pouvons affirmer que quel que soit le pas de temps $t \in [1, T]$, les cashflows associés au portefeuille agrégé présentent une erreur relative inférieure à 1% par rapport aux cashflows du portefeuille initial. L'application de ce critère conduit à déterminer des individus représentatifs d'un nombre $K < N$. Enlever un individu représentatif de cette liste conduirait à ce que le critère ne soit plus respecté.

Pérennité de la méthode d'agrégation

Nous avons vu que si l'on dispose des cashflows associés aux individus, il est possible de représenter une base de données contenant initialement N individus par une base contenant uniquement $K < N$ individus représentatifs.
Mais disposer des cashflows associés aux individus suppose que les projections ont été

réalisées par l'outil de modélisation. Il n'y a donc à ce stade pas d'intérêt à cette étape d'agrégation seule, car les calculs que l'on souhaite s'abstenir de réaliser sont précisément réalisés pour mettre en œuvre l'agrégation.
L'idée n'est donc pas d'utiliser cette méthode pendant les calculs, mais seulement en amont et surtout en association avec un autre outil qui ne nécessitera pas de disposer des cash-flows associés à une liste d'individus pour leur associer leurs individus représentatifs.
Les méthodes d'apprentissage automatisé (i.e. de classification supervisée) fournissent une réponse intéressante pour construire un tel outil. Le principe consiste à entraîner un algorithme sur une base d'apprentissage adéquate, afin de réaliser des regroupements d'assurés sans que les cashflows futurs probables ne soient calculés.

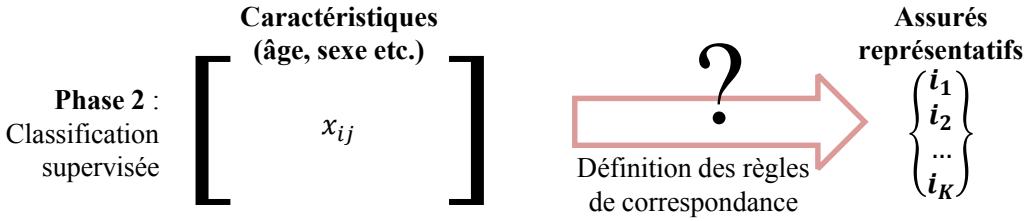


Schéma 7.3. Exploitation des méthodes de classification supervisée

Le développement de l'outil d'agrégation du passif se fait donc en deux temps :
- Dans un premier temps, la méthode des K-means n'est pas appliquée sur le portefeuille réel mais sur une base de données plus large, construite en opérant des perturbations sur les variables caractérisant les individus ;
- Dans un deuxième temps, l'outil d'agrégation est construit en utilisant un algorithme d'apprentissage automatisé.

Une fois l'outil d'agrégation construit, il peut être utilisé pour agréger le portefeuille à différentes dates d'arrêtés, tant que les hypothèses sous-jacentes n'ont pas été modifiées.
Les perturbations appliquées au portefeuille réel reproduisent les différentes variations que peut subir le passif, notamment le vieillissement des assurés ou les changements de leur état de sinistralité.
Le mémoire présente les méthodes d'apprentissage automatisé nécessaires à la mise en place de la méthode d'agrégation. Il s'agit principalement des méthodes des forêts aléatoires et du Boosting adaptatif.

Synthèse et analyse des résultats obtenus

Le schéma 7.4 donne une synthèse des différentes étapes de mise en place de la méthodologie proposée.

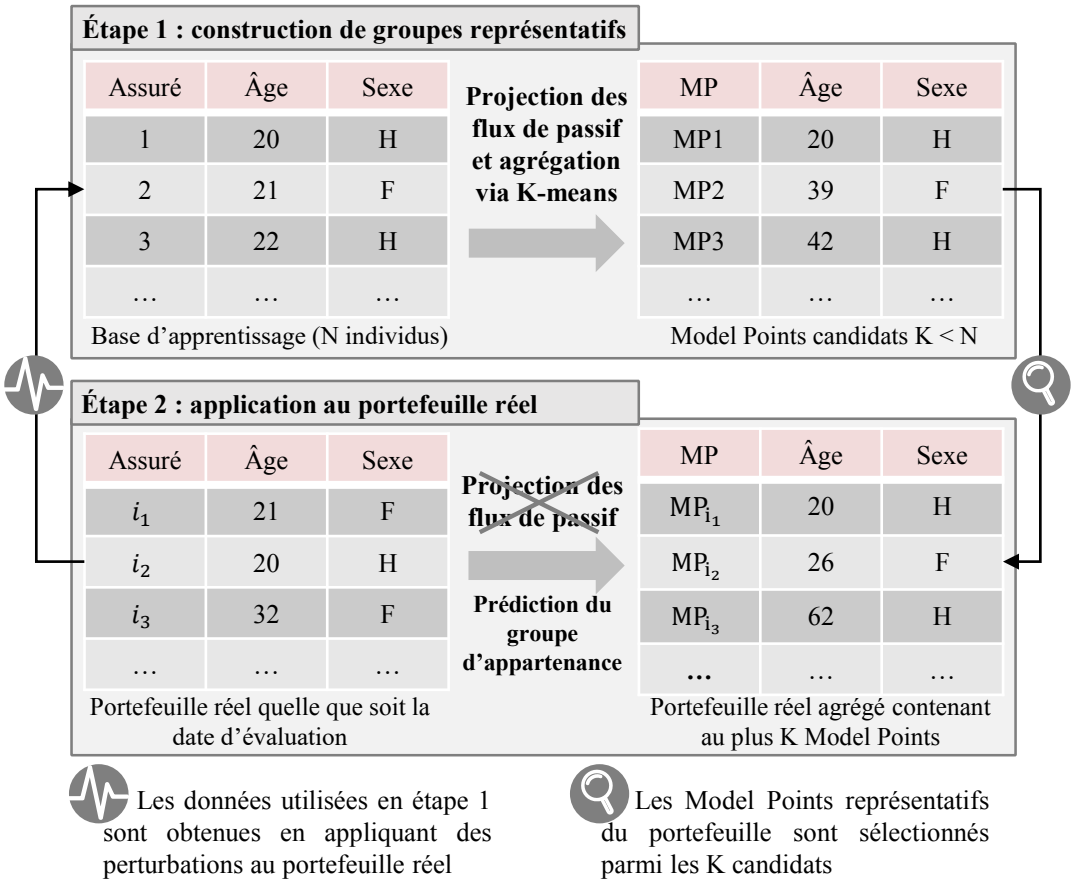


Schéma 7.4. Synthèse de la méthodologie proposée

En appliquant la méthode au portefeuille étudié, le nombre d'individus à projeter est considérablement réduit avec des impacts peu matériels sur le montant des provisions Best Estimate comme le montre le tableau ci-dessous :

	Best Estimate	Nombres de lignes projetées
Avant agrégation	2 390 517	66 212
Après agrégation	2 389 721	7 218
Erreur relative	0.03%	

Tableau 7.4. Impact de l'agrégation du passif sur les provisions Best Estimate

L'analyse des résultats fait ressortir les principales conclusions suivantes :

1– La taille du fichier d'inventaire du passif projeté est divisée par 9 pour une erreur relative très faible. À ce niveau d'erreur relative, la taille du fichier est nettement plus réduite qu'elle ne peut généralement l'être avec des méthodes classiques d'agrégation des passifs reposant sur les caractéristiques des assurés ;

2– L'efficacité de la méthode croît avec le nombre de paramètres entrant en jeu dans la modélisation de chaque garantie. Pour la garantie rente éducation, qui requiert le plus grand nombre de paramètres parmi les risques étudiés, la méthode permet de diviser par 50 le nombre d'individus à modéliser. Par ailleurs, l'analyse des regroupements réalisés fait ressortir une relative hétérogénéité au sein des groupes d'individus constitués : un même individu représentatif peut représenter des bénéficiaires homme ou femme, avec des niveaux de taux techniques différents ;

3– Le gain en temps de calcul est davantage visible lors de l'évaluation du capital réglementaire Solvabilité II car les chocs appliqués nécessitent de réévaluer les cashflows pour chaque scénario choqué.

L'impact de l'agrégation sur le ratio de solvabilité a été évalué en appliquant les différents chocs techniques et financiers de la formule standard. Le tableau ci-dessous présente cet impact qui s'avère très peu matériel.

	Avant agrégation	Après agrégation
Ratio de solvabilité	127.696%	127.692%

Tableau 7.5. Impact de l'agrégation sur le ratio de couverture Solvabilité II

Conclusion

En s'appuyant sur les algorithmes d'apprentissage automatisé, la méthodologie proposée est stable dans le temps, c'est-à-dire qu'elle est insensible au vieillissement du portefeuille et l'évolution du passif d'une année à l'autre. Son utilisation sur le portefeuille étudié fait ressortir des résultats satisfaisants, au regard de l'impact peu matériel qu'elle a sur les grandeurs d'intérêt et de la diminution substantielle d'informations relatives aux passifs induite par l'agrégation.

Le portefeuille étudié étant composé de garanties différentes, l'analyse détaillée des résultats permet par ailleurs de comparer les performances de la méthode suivant les risques concernés. Cet exercice permet d'illustrer le fait que l'efficacité de la méthode croit en fonction du degré d'élaboration de la modélisation. Cette propriété est particulièrement intéressante dans un contexte où l'amélioration de la qualité des données, motivée principalement par des raisons réglementaires ou commerciales, rend possible le développement d'une modélisation de plus en plus sophistiquée.

L'agrégation du passif offre beaucoup d'avantages d'un point de vue opérationnel. En outre, elle permet de réduire significativement les coûts liés aux plateformes informatiques nécessaires à la modélisation. C'est d'ailleurs un sujet d'actualité dans un contexte réglementaire plus complexe, notamment avec l'entrée en vigueur de la norme IFRS17.

Summary note

Context

Best Estimate Liabilities assessment requires to project the cashflows related to all the contracts underwritten by an insurance entity. Given the huge number of contracts held by big insurance compagnies, the separate calculation of the cashflows associated with each contract may be very time and memory consuming.

The guidelines (UE) 2015/35 issued by the European Commission regarding BEL calculation specify that insurers are allowed to use aggregation techniques for contracts that cover a similar risk, as long as the aggregation does not lead to a significant impact on the results.

This paper proposes an aggregation approach applied to an in-force business of income protection and health. However, it should be noted that the principles of the method can be extended to several life line of businesses, such as savings or loan insurance products.

This topic is of major interest because it allows to conduct easily subsequent studies such as ALM and ORSA requirements. In addition to this, using aggregation techniques contributes to reducing IT platforms' costs.

Several bibliographic resources deal with the topic of aggregating contracts. The existing methods rely on the characteristics of the policyholders (for instance, the variables age and sex) to create groups of similar policyholders. For example, if two pension beneficiaries have the same sex and were born the same year, they can be projected through a single representative individual.

The aim of this paper is to develop a new aggregation methodology : instead of using characteristics such as age and sex, the cashflows associated with policyholders are used to create representative groups. This will allow us to identify more interesting groups that those identified usually when relying on the characteristics of policyholders.

To understand this point, let us take the example of an annual term life insurance contract, based on TH/TF00-02 mortality tables, we can compute probabilities of death as shown in table 7.3.

Policyholder	Annual probability of death
A 46-years-old female	$0,1951\%$
A 38-years-old male	$0,1953\%$

Table 7.3. Illustrative example of similarity between policyholders

The two policyholders of the example have close probabilities of death and hence they can be put in the same cluster of policyholders. However, the two individuals do not present similarities based on age and sex.

Presentation of the in-force business

The main guarantees offered to policyholders or beneficiaries are summarized in the table below :

Guarantee	Indemnity covered
Temporary disability	Covering income daily benefits in case of a temporary disability
Permanent disability	Payment of a monthly pension in the case of permanent disability
Health	Reimbursement of health care costs
Death	Life term annual contracts (payment of a capital)
Death guarantee retention	Retention of death benefits following disability
Education annuity	Payment of an education pension in the event of the parent's death
Spouse annuity	Payment of a spouse's pension in the event of policyholder's death

Table 7.4. Risks and guarantees covered

The aggregation method applies to all guarantees except health and annual death, which are modelled through average profils.

For each beneficiary, we have variables such as age, sex or the date of occurrence of the claim. Those variables are required to project the cashflows of each contract by using two components :

- Firstly, technical assumptions (such as the probabilities of death or of remaining in disability state) ;
- Secondly, an actuarial modeling tool. The modeling approach used for our study is based on a state change model, in other words, the cashflows are obtained by projecting the number of policyholders in each state (disabled, dead and not affected policyholders).

The following table offers a summary of the parameters that have an impact on each guarantee as well as the underlying assumptions needed for cashflows projection.

Guarantee	Policyholders / beneficiaries characteristics	Technical assumptions
Temporary and permanent disability **Death guarantee retention**	- Age of the policyholder - Date of the claim's occurrence	- Probability of death or of remaining in disability state, identical for male/female
Education annuity	- Age of the beneficiary - Socio-professional category of the parent - Technical rate - Frequency of the pension	- Mortality table TGH/TGF-05 - Probability of education pursuit (by socio-professional category)
Spouse annuity	- Age of the beneficiary - Technical rate - Frequency of the pension	- Mortality table TGH/TGF-05

Table 7.5. Characteristics and parameters by guarantee

The aim of the method is to reduce the number of policyholders and beneficiaries projected using the actuarial modeling tool, while insuring that the impact on the results remains non significant.

The process of creating representative individuals

The data base used in this section contains characteristics of the policyholders as well as their future cashflows.

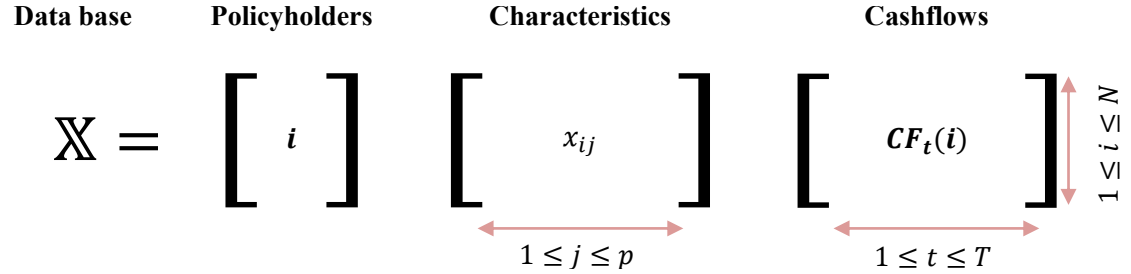


Figure 7.3. Nature of used data and notations

The choice of the adequate aggregation cashflows depends on the considered guarantee and on the assumptions impacting the projection. Given the parameters and the underlying assumptions used in this study, the cashflows used for aggregation are generally the future claims or the mathematical reserves.

The first step of our aggregation method is to identify individuals able to represent the real policyholders.

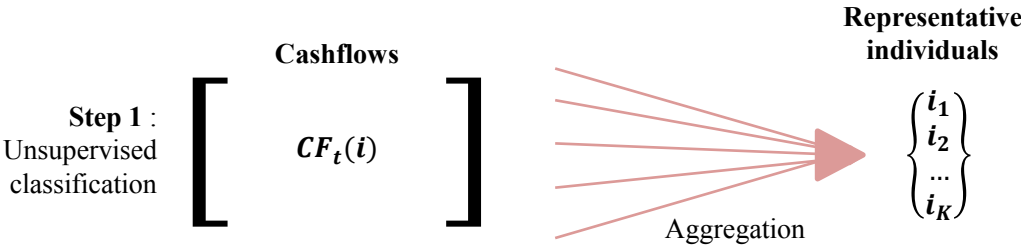


Figure 7.4. Presentation of the usage of unsupervised classification methods

The process of creating representative individuals is based on **unsupervised classification methods and more precisely on the K-means**.

The application of the K-means method requires to define several parameters. These are presented in detail in the paper, however, it is important to define the criterion used to validate the aggregation and to determine the number of representative groups.

At each projection step t, the cashflows of the liabilities affect the profit-sharing calculation and the Asset-Liability interaction. Therefore, the criterion to be defined must take into account the impacts of the aggregation at each time step.
Let us note, for a time step t, CF_t the sum of the cashflows of a group of policyholders before aggregation and CF_t^{ag} the cashflows obtained after the aggregation of the database. The criterion used to validate the aggregation is as follows :

$$\max_{t\in[1,T]} \left| \frac{CF_t - CF_t^{ag}}{CF_t} \right| < 1\% \tag{7.3}$$

Based on this criterion, we can say that for each time step $t \in [1, T]$, the relative error between cashflows before and after aggregation is lower than 1%.
The number of representative groups is obtained by increasing the value of this parameter of the K-means until our criterion is verified.

Finally, each group created is represented by the policyholder whose cashflows are the closest to those of the group's barycenter. In other words, a representative group is transformed to a representative individual.

Permanence of the aggregation method

We have seen that if we have cashflows associated with policyholders, it is possible to reduce a database containing N policyholders to only $K < N$ representative individuals. An important question arises at this stage, if we already have cashflows associated with the policyholders, the aggregation method would have minor interest because it would first require to project the real portfolio with our modeling tool.

Supervised classification algorithms provide an interesting answer to this issue. In other words, we can apply K-means on a training database, instead of the real portfolio, in order to create groups of policyholders.

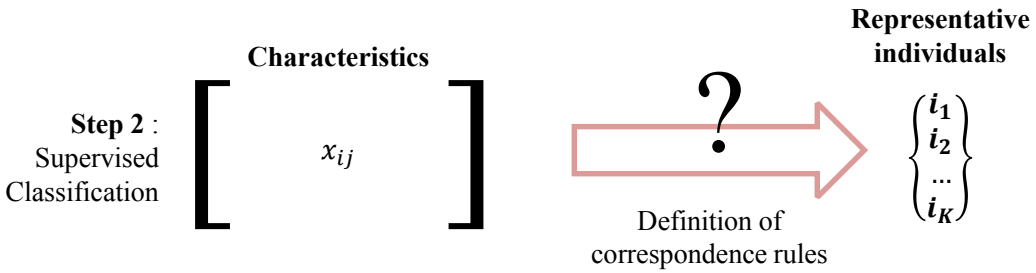


Figure 7.5. Presentation of the usage of supervised classification methods

As a result, the proposed aggregation method is realized following a two-steps approach :

- Firstly, rather than applying the K-means method on the real portfolio, we apply it on a new over-sampled training database, containing perturbed values of the real data ;
- Secondly, the representative group of each real policyholder is identified using a machine learning algorithm.

This method enables to aggregate the portfolio at different closing processes, as long as the underlying assumptions have not been modified.

The over-sampling of the real portfolio reproduces the variations that the contracts may undergo, particularly the aging of the portfolio and the changes of states of policyholders. In brief, our training data mirrors the real portfolio.

Furthermore, the main principles of supervised classification methods are presented in the paper, particularly random forests and adaptive boosting.

Application of the method and results analysis

The figure 7.6 gives an overall insight on the different stages of implementation of the aggregation method.

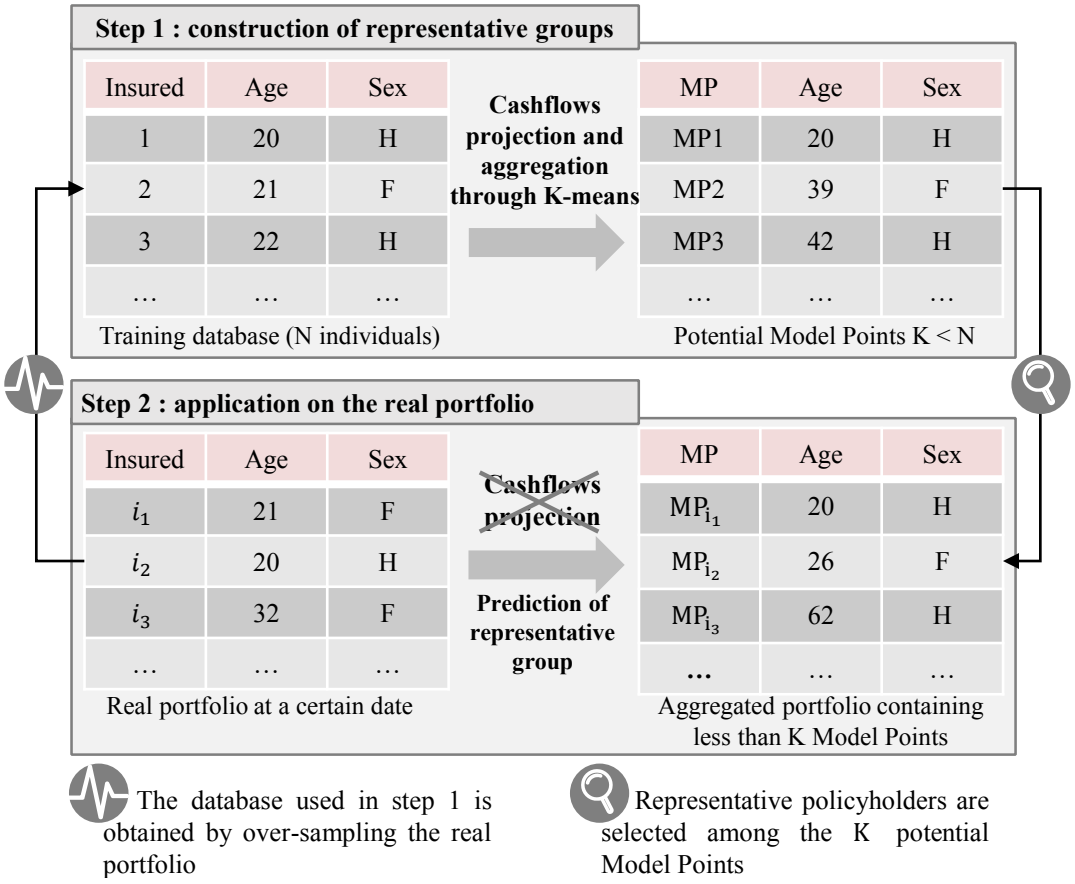


Figure 7.6. Summary of the aggregation method

As a result of the application of our method to the real portfolio, the number of Model Points to be projected is considerably reduced with no significant impact on the Best Estimate Liabilities :

	Best Estimate	Number of Model Points
Before Aggregation	2 390 517	66 212
After Aggregation	2 389 721	7 218
Relative error	0.03%	

Table 7.6. Impact of the aggregation on Best Estimate Liabilities

An analysis was carried out on the results following the application of the aggregation method. A particular mention should be made on the following conclusions :

1– The size of the projected data base is divided by 9. This reduction rate is much higher than those obtained using common methods (based exclusively on the characteristics of the policyholders) ;

2– The efficiency of the method increases with the number of parameters involved in the modeling of each guarantee. For instance, it was possible to divide by 50 the number of Model Points to be projected on education annuity's scope. In addition to this, the analysis of the realized groupings shows that there is heterogeneity within the groups created : we have within the same group male and female beneficiaries, with different technical rates ;

3– The impact on the run-time is more consistent when calculating Solvency Capital Requirement (SCR) because the shocks to be applied require the re-calculation of the cashflows of each scenario.

The impact on the solvency ratio is also assessed by applying life and market shocks defined in the standard formula. The results obtained are presented below :

	Before aggregation	After aggregation
Solvency Ratio	127.696%	127.692%

Table 7.7. Impact of the aggregation of the Solvency Ratio

Conclusion

The aggregation methodology allows to reduce significantly the number of Model Points to be projected, with a non-significant impact on Best Estimate Liabilities neither on the Solvency Ratio. In addition to this, by using machine learning algorithms, it was possible to make our method stable over time.

Furthermore, it is important to note that the implementation of this method as part of the production of regulatory reports requires several controls on financial indicators and regular back-testing.

Finally, the results of this paper can be used more broadly when it comes to assessing the Best Estimate Liabilities of a very large number of contracts, for example for other reporting and financial standards such as IFRS17.