

**Mémoire présenté le : 12/07/2022**

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

Par : Cochard Hugo

Titre : Utilisation du *machine learning* dans la détermination d'une allocation optimale sous Solvabilité II

Confidentialité :  NON  OUI (Durée :  1 an  2 ans)

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

*Membre présents du jury de l'Institut des Actuaires*

signature

*Entreprise : Optimind*

Bordelet Sophie


Nom :

Signature :

*Membres présents du jury de l'ISFA*

*Directeur de mémoire en entreprise :*

*Nom : Ngomo Leonel*

*Signature : *

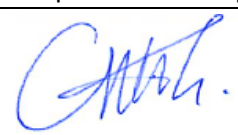
*Invité :*

Nom :

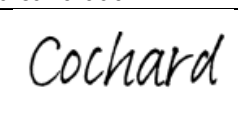
Signature :

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

Signature du responsable entreprise



Signature du candidat



# Résumé

Chez les assureurs, le placement de l'argent obtenu des souscripteurs est un élément essentiel car c'est ce mécanisme qui va permettre de générer des flux financiers. De plus, ces derniers doivent restituer les primes acquises sans perte si le risque est avéré, ainsi faire le bon choix de la stratégie d'allocation d'actifs est primordial, parmi la multitude d'options de placements possibles. Dans la pratique, l'identification d'une allocation optimale passe bien souvent par la projection d'un bilan sous Solvabilité II, avec un modèle de gestion actif-passif, afin de déterminer différents indicateurs de solvabilité et de rendement. Généralement, la méthode employée consiste à calculer des indicateurs tel que le SCR, pour un grand nombre d'allocations prédéfinies. C'est une évaluation dont la complexité entraîne de longs temps de calcul, et qui est néanmoins nécessaire car l'assureur doit prendre en compte différents indicateurs de solvabilité pour respecter la réglementation, mais aussi de rendement.

Ce mémoire met en avant l'utilisation du *machine learning* dans ce cadre. Si des modèles de *machine learning* permettent de déterminer assez précisément les indicateurs de risque et de rendement retenus pour une allocation donnée, leur vitesse de calcul pourrait permettre un important gain de temps. L'objectif de ce mémoire est donc d'évaluer la pertinence de l'utilisation du *machine learning* comme méthode alternative à celles existantes pour déterminer l'allocation d'actifs d'un assureur, dans le cadre de Solvabilité II.

Pour ce faire, une base de données est construite à partir de la projection d'un portefeuille d'épargne avec un modèle ALM, sur un ensemble de points d'allocation. Cette base comprend notamment différents indicateurs de risque et de rendement obtenus à l'issue de la projection pour chacun des scénarios d'allocation. Les indicateurs retenus sont le SCR de marché pour le risque, et la PVFP monde réel de la part des fonds en euros du portefeuille pour le rendement. Des modèles GLM, *random forest* et XGboost sont ensuite construits et calibrés sur cette base afin de chercher à prédire ces indicateurs. La pertinence des modèles est ensuite testée sur une base de données issue du vieillissement du portefeuille sur 1 an sans intégration de nouveaux contrats. Une dernière base issue d'un second vieillissement intégrant cette fois de nouveaux contrats est construite, cela permet d'observer l'efficacité des modèles dans un contexte différent. Les travaux mettent en avant une bonne capacité de prédiction du *machine learning*, malgré un biais qui peut apparaître lorsque de nouveaux contrats sont intégrés à la suite du vieillissement. Les modèles de *random forest* se démarquent particulièrement, et montrent une meilleure performance que les GLM, pourtant bien souvent utilisées en assurance en raison de leur explicabilité. Finalement, l'utilisation du *machine learning* en complément d'une méthode plus classique s'avère efficace afin de présélectionner rapidement un ensemble de points d'allocation d'actifs, et apporte donc un rapport gain/coût indéniable pour un assureur.

---

Mots-clés : *machine learning*, SCR, PVFP, Solvabilité II, Allocation stratégique, ALM, GSE, Assurance-vie

---

# Abstract

For insurers, the investment of money obtained from underwriters is an essential element because this is the mechanism that will generate financial flows. Moreover, they must return the premiums earned without loss if the risk is proven, so making the right choice of the asset allocation strategy is essential, among the multitude of possible investment options. In practice, the identification of an optimal allocation often involves the projection of a balance sheet under Solvency II, with an asset-liability management model, in order to determine solvency and return indicators. Generally, the method used consists of calculating indicators such as the SCR, for a large number of predefined allocations. This is a complex evaluation that takes a long time to compute but is also necessary because the insurer must consider different indicators of solvency to comply with regulations, and at the same time, different indicators of return.

This paper highlights the use of machine learning in this context. If machine learning models can be used to determine fairly accurately the risk and return indicators used for a given allocation, their computational speed could save a lot of time. The objective of this thesis is therefore to evaluate the relevance of using machine learning as an alternative method to the existing ones in order to determine the asset allocation of an insurer, in the framework of Solvency II.

To do this, a database is built from the projection of a savings portfolio with an ALM model, on a set of allocation points. This database includes various risk and return indicators obtained from the projection of each allocation scenario. The indicators used are the market module of SCR for risk, and the real world PVFP of the portfolio's share of euro funds for return. GLM, random forest and XGboost models are then built and calibrated on this basis to try to predict these indicators. The relevance of the models is then tested on a database resulting from the aging of the portfolio over 1 year without integrating new contracts. Another database from a second aging process, this time integrating new contracts, is built, which allows to observe the efficiency of the models in a different context. The study highlights the good predictive capacity of machine learning, despite a bias that may appear when new contracts are integrated following the aging. The random forest models stand out and show a better performance than GLMs, which are often used in insurance because of their explicability. Finally, the use of machine learning as a complement to a more classical method proves to be efficient in order to quickly preselect a set of asset allocation points, and thus brings an undeniable gain/cost ratio for an insurer.

---

Key words : machine learning, SCR, PVFP, Solvency II, strategic allocation, ALM, ESG, life insurance

---

# Remerciements

Je souhaite remercier toutes les personnes ayant contribué à la réalisation de ce mémoire et qui se sont rendues disponibles malgré le contexte particulier de la crise sanitaire.

Tout d'abord je tiens à remercier Christophe Eberlé, Président et fondateur d'Optimind pour m'avoir permis d'effectuer mon mémoire au sein de son entreprise.

Je remercie aussi mon tuteur à Optimind Leonel Ngomo, qui a pu m'accompagner pour les travaux de ce mémoire. Son encadrement m'a permis d'augmenter grandement mes compétences dans les domaines étudiés.

J'adresse également mes remerciements à Hugo Tambrun, Andre Grondin, Jorge Ochoa, Nicolas Lorin, et Marc-Antoine Cottignies pour l'aide qu'ils m'ont apporté dans différentes parties du mémoire.

Merci à Clara Kibler, Bisous Razafindrabary, Marius Masson, William Fouques, Benjamin Aubignat, Jules Nahon, Lea Karim pour les bons moments que nous avons pu partager.

Plus généralement je remercie aussi toute l'équipe d'Optimind dont j'ai beaucoup apprécié l'accueil.

Finalement, je remercie chaleureusement ma famille et mon entourage pour leur soutien depuis tant d'années.

# Table des matières

<b>Résumé</b> .....	<b>i</b>
<b>Abstract</b> .....	<b>ii</b>
<b>Remerciements</b> .....	<b>iii</b>
<b>Introduction</b> .....	<b>1</b>
<b>I. Contexte général</b> .....	<b>2</b>
1.1 Introduction aux problématiques de rendement .....	2
1.2 Frontière efficiente de Markowitz .....	3
1.2.1 Présentation générale .....	3
1.2.2 Cadre du problème.....	3
1.2.3 Ajout de l'actif sans risque .....	6
1.2.4 Développements et limites .....	8
1.3 Expression du rendement par la PVFP .....	11
1.4 Le SCR de marché .....	13
1.4.1 Contexte général de la réglementation Solvabilité II .....	13
1.4.2 Calcul du SCR de marché .....	17
<b>II. Approche du problème</b> .....	<b>24</b>
2.1 Mise en avant de solutions d'allocation d'actifs.....	24
2.1.1 Méthode de descente du gradient.....	24
2.1.2 Algorithme génétique.....	25
2.1.3 Analyse critique .....	27
2.2 Focus sur le <i>machine learning</i> .....	28
2.3 Apprentissage non-supervisé et supervisé .....	30
2.4 Présentation des différents modèles .....	32
2.4.1 Modèle linéaire généralisé (GLM) .....	32
2.4.2 Arbres de décision .....	34
2.4.3 Forêts aléatoires.....	37
2.4.4 Extreme gradient boosting (XGboost).....	38
2.5 Généralités d'un contrat d'épargne, modèle ALM et GSE utilisés .....	41
2.5.1 Généralités du contrat d'épargne .....	41
2.5.2 Modèle ALM .....	43
2.5.3 Générateur de scénarios économiques (GSE) .....	49
<b>III. Construction de la base de calcul</b> .....	<b>55</b>
3.1 Présentation du portefeuille en situation initiale et hypothèses générales.....	55

3.2	Génération de la base de données.....	60
3.3	Vieillessement du portefeuille sur 1 an.....	64
<b>IV.</b>	<b>Utilisation du <i>machine learning</i> pour déterminer l'allocation optimale.....</b>	<b>66</b>
4.1	GLM .....	66
4.2	Random Forest .....	72
4.3	XGboost .....	80
4.4	Comparaison des modèles .....	86
	<b>Conclusion .....</b>	<b>88</b>
	<b>Lexique .....</b>	<b>A</b>
	<b>Liste des figures.....</b>	<b>B</b>
	<b>Liste des tableaux.....</b>	<b>D</b>
	<b>Bibliographie.....</b>	<b>E</b>

# Introduction

Lorsqu'il perçoit des primes, l'assureur va chercher à placer cet argent sur des actifs financiers. C'est par ce biais que son activité génère un profit. Ce dernier montre ainsi une volonté forte de maîtriser son allocation d'actifs, car c'est de celle-ci que va dépendre la génération des produits financiers. Par ailleurs, la détermination de sa stratégie d'allocation sous la réglementation Solvabilité II, passe par la projection d'un bilan avec un modèle de gestion actif-passif. C'est une évaluation conséquente dont la complexité entraîne un long temps de calcul, qui est néanmoins nécessaire car l'assureur doit prendre en compte différents indicateurs de solvabilité, mais aussi de rendement. La méthode généralement utilisée afin de déterminer l'allocation optimale des actifs pour l'assureur passe ainsi par le calcul d'indicateurs tel que le SCR, pour un grand nombre d'allocations prédéfinies. Cette démarche implique de nombreuses heures de calcul non productives et qui s'avèrent très coûteuses pour l'entreprise.

Ici se trouve l'intérêt de l'utilisation du *machine learning* dans ce cadre ; la construction de tels modèles est coûteuse, néanmoins si le *machine learning* permet de déterminer assez précisément des indicateurs de risque et de rendement pour une allocation donnée, le gain de temps pour l'évaluation de diverses allocations sous Solvabilité II pourrait être conséquent. L'objectif de ce mémoire est donc d'évaluer la pertinence de l'utilisation du *machine learning* comme méthode alternative à celles existantes pour déterminer l'allocation d'actifs d'un assureur, dans le cadre de Solvabilité II.

La démarche de réalisation de ce mémoire se fera ainsi de la façon suivante. Tout d'abord, un portefeuille d'épargne fictif constitué à dire d'experts et représentatif du marché français sera utilisé afin de créer des *models points* et de déterminer différentes hypothèses. La projection de ce portefeuille sous la réglementation Solvabilité II, avec un modèle ALM, et sur un ensemble de points d'allocations, nous permettra de générer une base de données initiale en  $T_0$ . Les métriques utilisées afin d'évaluer chacune des allocations testées seront le SCR de marché et la PVFP monde réel de la part des fonds en euros du portefeuille.

Un second travail se fera alors sur la construction de modèles de *machine learning* à partir de cette base de données pour déterminer ces deux indicateurs. Un vieillissement du portefeuille sera alors effectué sur une année, avant de tester à nouveau l'efficacité des modèles sur de nouvelles allocations en  $T_1$ .

L'objectif est donc d'être capable, via ces résultats sur un échantillon de points d'apprentissage, de construire un modèle permettant de prédire le plus précisément possible nos indicateurs, à un instant  $T_1$ , sans avoir besoin de réaliser à nouveau une projection avec un modèle ALM.

La suite de ce mémoire présentera le contexte de ces travaux ainsi que l'approche du problème.

# Contexte général

## 1.1 Introduction aux problématiques de rendement

La crise économique de 2008 a marqué le début des mesures prises par la Banque Centrale Européenne qui ont établi un environnement de taux bas en Europe. Depuis quelques années nous avons atteint des taux négatifs, les obligations étant auparavant des actifs financiers très rentables, se trouvent désormais peu rémunératrices. L'exemple peut être pris du taux obligataire allemand de maturité 10 ans qui avait un rendement supérieur à 2% avant 2012 et qui a chuté drastiquement jusqu'à atteindre -0,7% lors de l'été 2019.

L'évolution de ce contexte économique impacte fortement les entreprises dont le métier implique de détenir des actifs financiers. Les obligations représentaient historiquement une très importante voire une majeure partie de ces portefeuilles. Néanmoins, ces stratégies d'investissement impliquaient un rendement intéressant que n'ont plus les obligations. Il devient ainsi primordial pour ces entreprises d'adapter la composition de leur portefeuille.

De plus, chez les assureurs une arrivée à maturité des obligations acquises avant la crise est observée. Les réinvestissements se faisant désormais sur ce support avec des taux de coupons très bas, le produit financier se voit grandement diminué.

Ces problématiques impactent fortement le métier des banques et des assureurs. Le choix de l'allocation d'actifs financiers est la base du rendement qui sera obtenu par l'entreprise et est déterminant pour la pérennité du métier d'assureur. Néanmoins il faut associer à ce rendement la notion de risque qui l'accompagne car l'activité d'assurance est régulée. En effet, la nature du métier de l'assureur entraîne un cycle de production inversé, soit l'encaissement de primes aujourd'hui pour un versement en prestation demain. Ainsi il est primordial de maîtriser les risques auxquels est exposé l'assureur afin d'être en mesure de résister aux différents aléas pouvant arriver dans la vie d'un contrat.

L'optimisation financière doit donc être associée à une forte gestion des risques. La réglementation Solvabilité II mise en application depuis 2016 apporte un cadre au niveau Européen qui impose des risques contrôlés. L'enjeu du choix de l'allocation des actifs amène ainsi l'optimisation du couple rendement et risque lié au portefeuille. Plusieurs théories existent afin de proposer une approche à cette optimisation, c'est notamment le cas de la théorie moderne du portefeuille de Markowitz développée en 1952, qui pose les fondements du lien entre le risque et le rendement.



## 1.2 Frontière efficiente de Markowitz

### 1.2.1 Présentation générale

La théorie moderne du portefeuille de Markowitz pose la problématique suivante : quelle est la composition optimale d'un portefeuille ? Il définit ainsi la frontière efficiente comme l'ensemble des portefeuilles optimaux efficients, c'est-à-dire, l'ensemble des portefeuilles qui, pour tout risque donné, vont maximiser le rendement et pour tout rendement donné vont minimiser le risque. Le couple rendement-risque correspond ici à l'espérance et la variance de la combinaison linéaire pondérée d'actifs susceptibles de faire partie du portefeuille. L'indépendance entre les différents actifs n'étant jamais parfaite, le facteur de corrélation entre les différents actifs est ensuite pris en compte. D'après Markowitz, la diversification des actifs permet une forte réduction de la variance du portefeuille et donc d'obtenir un portefeuille efficient en son sens.

### 1.2.2 Cadre du problème

On suppose que le choix de l'investisseur pour un couple risque-rendement donné peut être traduit par une fonction d'utilité quadratique dont on cherche à maximiser l'espérance. Ses deux premiers moments sont :

$$\begin{aligned}u(W) &= a + bW + cW^2 \\ E[u(W)] &= a + bE[W] + c(Var(W) + E[W]^2) \\ &\forall b > 0, \forall c < 0\end{aligned}$$

De plus, il est supposé que les évolutions de marché suivent une loi normale, et donc les seuls paramètres retenus par l'investisseur sont l'espérance (le rendement) ainsi que la variance (le risque).

On observe le marché lors de la date de constitution du portefeuille en  $t = 0$ , ainsi qu'à la date de réalisation en  $t = 1$ . Si un portefeuille à  $n$  actifs risqués est considéré, le rendement aléatoire de l'actif  $i$  peut être défini par :

$$\tilde{R}_i = \frac{\tilde{P}_i(1) - P_i(0)}{P_i(0)}$$

Avec  $P_i(0)$  la valeur de l'actif en  $t = 0$ , et  $\tilde{P}_i(1)$  la valeur aléatoire de l'actif en  $t = 1$ .

Nous posons  $\omega_i$  le poids associé à chaque actif  $i$ . Le rendement aléatoire du portefeuille est déduit par :

$$\tilde{R}_p = \sum_i \omega_i \tilde{R}_i$$

Avec  $i = 1, \dots, n$

Le risque ou volatilité du portefeuille s'écrit  $\sigma_p$  avec  $\sigma_p^2 = Var(\tilde{R}_p)$  et on écrit  $\Sigma$  la matrice de variance-covariance de terme  $\sigma_{ij} = cov(\tilde{R}_i, \tilde{R}_j)$  avec  $\sigma_{ii} = \sigma_i^2 = Var(\tilde{R}_i)$ . On suppose que la matrice  $\Sigma$  existe et est inversible. Ainsi, la variance du portefeuille correspond à la somme des produits des poids de chaque couple d'actifs multipliés par leur covariance, soit :

$$\sigma_p^2 = \sum_i \sum_j \omega_i \omega_j \sigma_{ij}$$

Avec  $i = 1, \dots, n$  et  $j = 1, \dots, n$ .

Nous avons ainsi le rendement espéré du titre  $i$  de  $t = 0$  à  $t = 1$  :

$$\mu_i = E[\tilde{R}_i]$$

Dont il découle le rendement espéré du portefeuille :

$$\mu_p = E[\tilde{R}_p]$$

Enfin, il est possible d'écrire :

- le vecteur de rendement aléatoire des  $n$  actifs  $\tilde{R} = (\tilde{R}_1, \tilde{R}_2, \dots, \tilde{R}_n)^T$
- le vecteur de rendement espéré des  $n$  titres  $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$
- Le vecteur de poids associé aux  $n$  actifs, soit la composition du portefeuille,  $\omega = (\omega_1, \dots, \omega_n)^T$  tel que  $\sum_{i=1}^n \omega_i = 1$ .

Finalement, il est possible de simplifier certaines égalités en les mettant sous la forme de produit de vecteurs :

- Le rendement aléatoire du portefeuille  $\tilde{R}_p = \omega^T \tilde{R}$
- Le rendement espéré du portefeuille  $\mu_p = \omega^T \mu$
- La variance du portefeuille  $\sigma_p^2 = \omega^T \Sigma \omega$

Pour un rendement cible donné  $\mu_{obj}$ , un programme d'optimisation sous contrainte permettant d'obtenir la proportion optimale de chaque actif, soit celle minimisant la variance du portefeuille, peut ainsi être créé.

$$\begin{aligned} & \min_{\omega \in \mathbb{R}^n} \omega^T \Sigma \omega \\ & \text{s.c. } \sum_{i=1}^n \omega_i = 1 \\ & \text{s.c. } \omega^T \mu = \mu_{obj} \end{aligned}$$

Le Lagrangien du problème s'écrit :

$$L(\omega, \lambda_1, \lambda_2) = \omega^T \Sigma \omega - \lambda_1 \left( \sum_{i=1}^n \omega_i - 1 \right) - \lambda_2 (\omega^T \mu - \mu_{obj})$$

Nous en déduisons les dérivées partielles et les équations suivantes :

$$\begin{aligned} \frac{\partial L(\omega, \lambda_1, \lambda_2)}{\partial \omega} &= 2\Sigma\omega - \lambda_1 \mathbb{1}_{\mathbb{R}^n} - \lambda_2 \mu = 0 \\ \frac{\partial L(\omega, \lambda_1, \lambda_2)}{\partial \lambda_1} &= - \left( \sum_{i=1}^n \omega_i - 1 \right) = 0 \\ \frac{\partial L(\omega, \lambda_1, \lambda_2)}{\partial \lambda_2} &= -(\omega^T \mu - \mu_{obj}) = 0 \end{aligned}$$

On en déduit donc la relation suivante :

$$\omega = \frac{\lambda_1}{2} \Sigma^T \mathbb{1}_{\mathbb{R}^n} + \frac{\lambda_2}{2} \Sigma^T \mu$$

Pour résoudre le problème il faut vérifier le système d'équations suivant :

$$\begin{cases} \omega = \frac{\lambda_1}{2} \Sigma^T \mathbb{1}_{\mathbb{R}^n} + \frac{\lambda_2}{2} \Sigma^T \mu \\ - \left( \sum_{i=1}^n \omega_i - 1 \right) = 0 \\ - (\omega^T \mu - \mu_{obj}) = 0 \end{cases} \Leftrightarrow \begin{cases} \omega = \frac{\lambda_1}{2} \Sigma^T \mathbb{1}_{\mathbb{R}^n} + \frac{\lambda_2}{2} \Sigma^T \mu \\ \lambda_1 \mathbb{1}_{\mathbb{R}^n}^T \Sigma^T \mathbb{1}_{\mathbb{R}^n} + \lambda_2 \mathbb{1}_{\mathbb{R}^n}^T \Sigma^T \mu = 2 \\ \lambda_1 \mu^T \Sigma^T \mathbb{1}_{\mathbb{R}^n} + \lambda_2 \mu^T \Sigma^T \mu = 2\mu_{obj} \end{cases}$$

On pose les scalaires A, B et C tels que :

$$\begin{aligned} A &= \mathbb{1}_{\mathbb{R}^n}^T \Sigma^T \mathbb{1}_{\mathbb{R}^n} \\ B &= \mathbb{1}_{\mathbb{R}^n}^T \Sigma^T \mu \\ C &= \mu^T \Sigma^T \mu \end{aligned}$$

Nous pouvons à présent résoudre le système d'équations et obtenir après simplification :

$$\begin{aligned} \lambda_1 &= 2 \frac{(C - B\mu_{obj})}{CA - B^2} \\ \lambda_2 &= 2 \frac{(A\mu_{obj} - B)}{CA - B^2} \\ \omega^* &= \frac{(C - B\mu_{obj})}{CA - B^2} \Sigma^T \mathbb{1}_{\mathbb{R}^n} + \frac{(A\mu_{obj} - B)}{CA - B^2} \Sigma^T \mu \end{aligned}$$

Avec  $\omega^*$  le portefeuille de variance minimale, solution de notre problème d'optimisation. On doit désormais calculer la variance qui correspond au portefeuille optimal en fonction de  $\mu_{obj}$ .

$$\sigma_P^2 = \omega^{*T} \Sigma \omega^*$$

Car  $\omega^*$  vérifie les contraintes de notre problème d'optimisation. On obtient ainsi :

$$\begin{aligned} \sigma_P^2 &= \left( \frac{\lambda_1}{2} \Sigma^T \mathbb{1}_{\mathbb{R}^n} + \frac{\lambda_2}{2} \Sigma^T \mu \right)^T \Sigma \omega^* \\ \sigma_P^2 &= \frac{\lambda_1}{2} + \frac{\lambda_2}{2} \mu_{obj} \\ \sigma_P^2 &= \frac{(C - B\mu_{obj})}{CA - B^2} + \frac{(A\mu_{obj} - B)}{CA - B^2} \mu_{obj} \\ \sigma_P^2 &= \frac{1}{CA - B^2} (A\mu_{obj}^2 - 2B\mu_{obj} + C) \end{aligned}$$

En faisant varier  $\mu_{obj}$  la frontière efficiente peut ainsi être obtenue.

Illustrons cette Théorie dans le cas d'un portefeuille de 2 actifs ayant potentiellement les rendements suivants à la suite de la période observée :

A	B
-5%	12%
0%	8%
5%	9%
10%	-3%
15%	5%

Ces évènements sont supposés équiprobables. On peut calculer rapidement que  $\mu_A = 5\%$ ,  $\mu_B = 6.2\%$  et  $\sigma_A^2 = 0.005$ ,  $\sigma_B^2 = 0.002616$ .

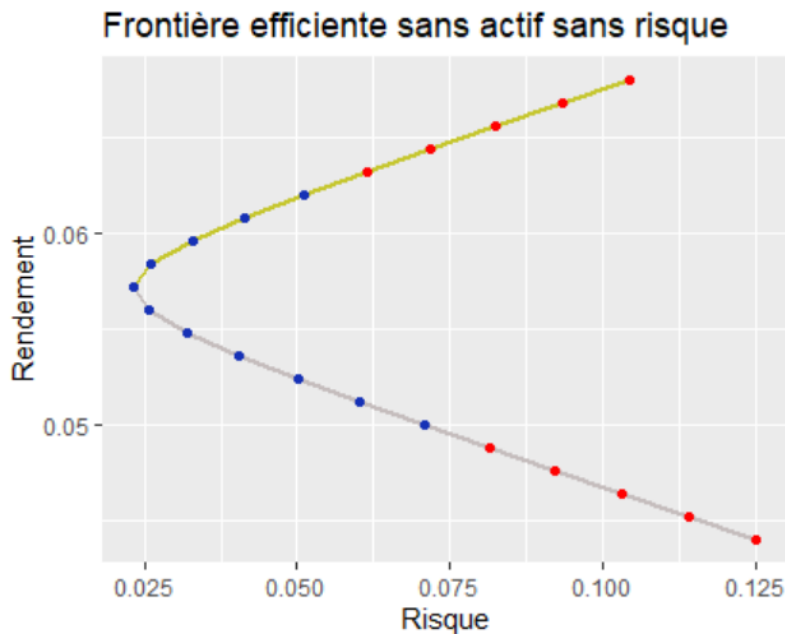


Figure 1 - Représentation d'une frontière efficiente

Il est possible d'observer sur le graphique ci-dessus la frontière efficiente du portefeuille en jaune. Les points rouges montrent les cas où la vente à découvert est autorisée. Ainsi, la partie de la courbe en gris n'est pas intéressante pour un investisseur rationnel car pour un même niveau de risque il peut obtenir un rendement espéré plus élevé. Il cherchera donc à se situer sur la courbe en jaune au niveau de risque qu'il aura choisi.

### 1.2.3 Ajout de l'actif sans risque

On appelle sans risque un actif théorique qui a pour rendement le taux d'intérêt sans risque et une variance nulle. Étant donné sa non-corrélation avec les autres actifs d'un portefeuille, il modifie de façon linéaire son rendement espéré ainsi que sa variance. Les taux d'intérêt sans risque correspondent en général aux taux d'emprunts d'états court terme.

Nous avons désormais pour l'actif sans risque  $f$  le rendement espéré du portefeuille qui vaut :

$$\mu_p = R_f + \sum_i \omega_i [\tilde{R}_i - R_f]$$

Et nous n'avons plus la contrainte  $\sum_{i=1}^n \omega_i = 1$ . Le modèle se pose à présent de la façon suivante :

$$\min_{\omega \in \mathbb{R}^n} \omega^T \Sigma \omega$$

$$\text{s.c. } \omega^T \mu + [1 - \sum_{i=1}^n \omega_i] \times R_f = \mu_{obj}$$

Le Lagrangien du problème s'écrit :

$$L(\omega, \lambda) = \omega^T \Sigma \omega - \lambda \left( \omega^T \mu + [1 - \sum_{i=1}^n \omega_i] \times R_f - \mu_{obj} \right)$$

En suivant la même méthode qu'avant l'inclusion de l'actif sans risque, nous obtenons donc :

$$\sigma_P = \begin{cases} \frac{\mu_{obj} - R_f}{\sqrt{\pi^T \Sigma^T \pi}}, & \mu_{obj} \geq R_f \\ \frac{R_f - \mu_{obj}}{\sqrt{\pi^T \Sigma^T \pi}}, & \mu_{obj} < R_f \end{cases}$$

Avec  $\pi = \mu - R_f \mathbb{1}_{\mathbb{R}^n}$ , le vecteur de prime de risque.

En prenant un portefeuille composé d'un actif risqué A et d'un actif sans risque B de rendement  $R_f = 3\%$ , soit un portefeuille avec les rendements suivants pour A et B non corrélés.

A	B
-5%	3%
0%	3%
5%	3%
10%	3%
15%	3%

On peut désormais tracer la frontière efficiente du portefeuille.

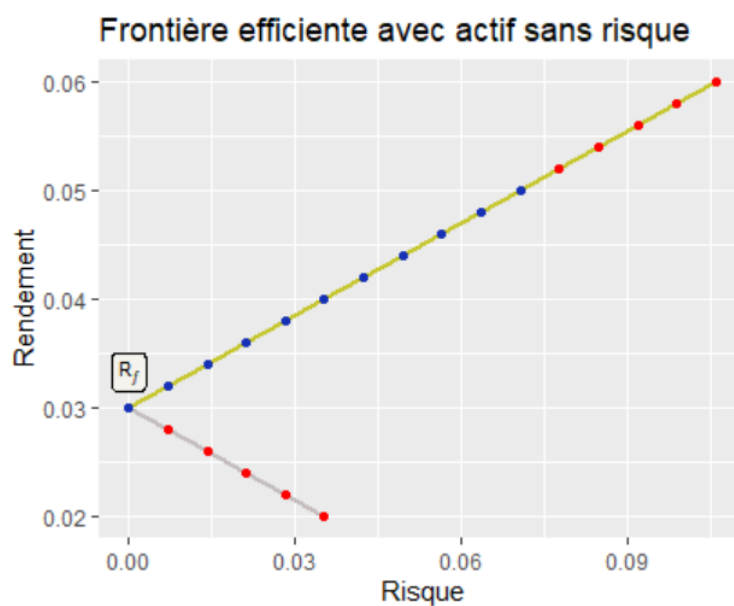


Figure 2 - Représentation d'une frontière efficiente après intégration de l'actif sans risque

L'effet linéaire de l'ajout de l'actif sans risque est très visible. La frontière efficiente est donc dans ce cas une demi-droite avec en extrémité un risque nul pour un portefeuille comprenant exclusivement l'actif sans risque. On remarque aussi que les positions courtes sur l'actif sans risque ne peuvent pas faire partie d'un portefeuille efficient.

### 1.2.4 Développements et limites

La théorie de Markowitz permet donc une première approche aux problèmes d'allocation d'actifs, néanmoins certaines limites rendent difficile l'utilisation de ce modèle dans un cadre plus pratique. Les modèles suivants ont ainsi été développés.

#### Modèle d'évaluation des actifs financiers (MEDAF)

Cette méthode fait suite aux travaux d'Harry Markowitz et est introduite pour la première fois en 1961. Elle explique que l'offre et la demande d'un titre financier va permettre de définir un équilibre de marché. Le MEDAF implique le traçage de la demi-droite de marché des capitaux (CML) en fonction du risque et du rendement espéré. Cette demi-droite part de  $R_f$  et est tangente à la frontière efficiente. L'investisseur pourra alors choisir un portefeuille entre le point  $R_f$  et le point tangent entre la CML et la frontière efficiente. L'idée est donc que la rentabilité espérée et le risque de tout actif sont liés par une relation linéaire. La rentabilité espérée d'un titre  $i$  est donc la suivante :

$$\mu_i = R_f + \beta_i(\mu_M - R_f)$$

Avec  $\beta_i = \frac{cov(\tilde{R}_i, \tilde{R}_M)}{var(\tilde{R}_M)}$  qui correspond à l'élasticité du titre par rapport à l'indice boursier représentant le marché.

Le MEDAF permet enfin de tracer la CML au niveau des titres individuels, qui est appelé la SML. Cela permet de représenter graphiquement la rentabilité d'un titre par rapport à son évaluation sur le marché. Ainsi, les titres situés au-dessus de la SML sont considérés comme pas assez cher, et donc sont un choix intéressant pour un investisseur alors que les titres situés au-dessous de la SML sont quant à eux surévalués.

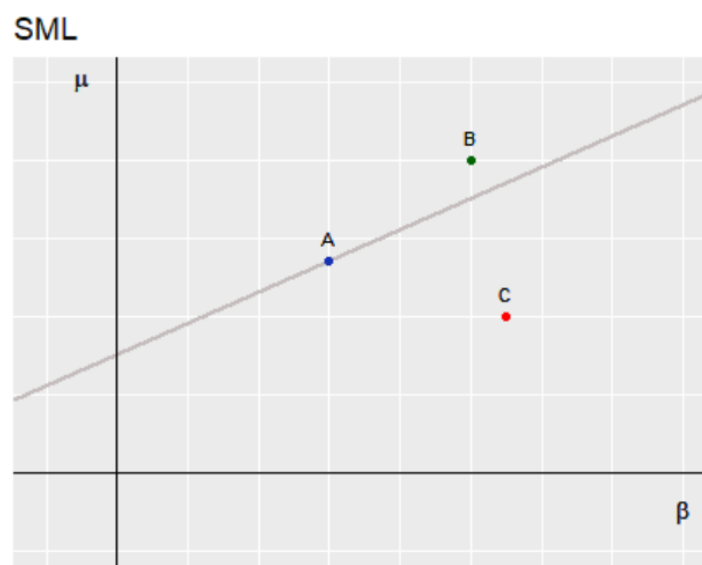


Figure 3 - Représentation de la CML avec 3 titres fictifs

Le titre B est ici une opportunité pour un investisseur car sa rentabilité est très forte par rapport à son risque. Le titre C n'est quant à lui pas assez rentable par rapport à la SML, et l'actif A est correctement évalué sur le marché.

### Modèle de Black-Litterman

Développé en 1992, ce modèle s'appuie sur la théorie de Markowitz et le MEDAF tout en cherchant à introduire les prévisions des investisseurs. Le rendement  $\mu$  devient maintenant aléatoire  $\tilde{\mu}$  et dépendra aussi de la confiance de l'investisseur en ses prévisions. Il faut ainsi définir un portefeuille de vues  $p$ , qui peuvent être absolues ou relatives, et y associer un vecteur de rendement  $q$  et une matrice d'incertitude  $\Omega$ . Prenons l'exemple d'un investisseur qui aurait 3 vues sur 6 actifs du marché :

- L'actif 1 aura un rendement de 10%, avec une confiance dans la vue de 50%.
- L'actif 2 surperformera l'actif 3 de 5%, avec une confiance dans la vue de 30%.
- L'actif 4 surperformera la somme des actifs 5 et 6 de 3%, avec une confiance dans la vue de 70%.

Nous avons donc :

$$p = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 & 1/2 \end{pmatrix}$$

$$q = \begin{pmatrix} 10\% \\ 5\% \\ 3\% \end{pmatrix}$$

$$\Omega = \begin{pmatrix} 50\% & 0 & 0 \\ 0 & 70\% & 0 \\ 0 & 0 & 30\% \end{pmatrix}$$

L'allocation des actifs sera ensuite faite de la même façon que dans la théorie de Markowitz, mais en prenant en compte le sentiment de l'investisseur à travers les paramètres ci-dessus.

### Limites

Malgré le développement de nombreux modèles faisant suite à la théorie de Markowitz, il reste de nombreuses limites à ces approches.

Tout d'abord l'unique période d'observation du marché entre  $t = 0$  et  $t = 1$  ne permet pas de refléter la stratégie d'un investisseur. Dans la pratique, ces derniers ont en général différents placements à horizon court, moyen et long terme et les modèles mono-période ne peuvent pas représenter ces stratégies. De plus, si une unique période à horizon lointain est considérée, il est difficile d'estimer un rendement et une covariance pour un actif, et la qualité du modèle dépendra fortement de ces paramètres.

En effet, l'estimation des paramètres est un point central de la Théorie de Markowitz car une légère variation de ces derniers peut entraîner des changements disproportionnés dans la composition de portefeuilles efficients. Une augmentation de quelques points de variance pour un actif pourrait par exemple porter son poids à 0% dans une allocation alors qu'il était utilisé précédemment. Le portefeuille de rendement cible pourrait donc voir sa composition varier au fil du temps car son rendement aléatoire en fin de période et sa variance sont susceptibles d'évoluer.

Le problème d'optimisation peut parfois proposer des résultats aberrants et difficilement applicables. Il peut notamment y avoir des allocations avec des positions nulles sur la plupart des actifs sauf quelques-uns, ou bien inclure d'importantes ventes à découvert car il n'y a pas de contrainte à ce sujet dans le modèle moyenne-variance de Markowitz. L'approche du MEDAF permet d'éviter en partie ces problématiques.

La résolution du problème d'optimisation implique aussi que la matrice de variance-covariance  $\Sigma$  est inversible. Dans la réalité ce n'est pas forcément le cas.

L'hypothèse de marché parfait du modèle de Markowitz n'est pas réaliste, les actifs ne sont que très rarement disponibles à leur juste prix. Le MEDAF permet de s'en rendre compte et de palier à cela avec la prise en compte du coefficient bêta, ou la représentation graphique de la CML. Néanmoins, il faut aussi prendre en compte l'instabilité du coefficient bêta dans le temps et la difficulté à observer le portefeuille de marché qui est souvent représenté par un indice comme le S&P 500.

Markowitz et le MEDAF supposent l'investisseur indépendant émotionnellement du marché, or au contraire celui-ci ne peut jamais être totalement objectif et trouve un biais dans sa perception du marché. Black-Litterman apporte ainsi une solution à ce problème afin de prendre en compte les sentiments de l'investisseur par la prise en compte de ses vues sur le marché ainsi que de sa confiance en ces vues.

Notre première hypothèse est que les évolutions de marché suivent une loi normale et que les investisseurs prennent des décisions en fonction de l'espérance et de la variance. Dans la réalité, la fonction d'utilité ne suit pas une loi normale et en l'occurrence cette loi sous-estime grandement les événements extrêmes. Pourtant ces événements ont un fort impact sur un portefeuille et les krachs boursiers, par exemple, sont finalement beaucoup moins rares.

Enfin, la mesure de risque choisi est la variance, néanmoins celle-ci représente une dispersion centrée sur la moyenne. Le risque pour un investisseur est que le rendement de son portefeuille se trouve sous son espérance, il serait donc plus judicieux de considérer un indicateur de baisse de rendement uniquement.

Dans le cadre de ce mémoire, les travaux seront effectués sous le contexte de Solvabilité II. La notion de frontière efficiente comme on a pu le voir est difficilement applicable en prenant comme mesure de performance l'espérance du rendement, et comme mesure de risque la variance. La métrique qui sera d'intérêt afin de mesurer le risque sera donc le SCR de marché, à laquelle sera associée comme mesure de performance la PVFP (*Present Value of Future Profit*) monde réel.



## 1.3 Expression du rendement par la PVFP

La PVFP, *Present Value of Future Profit*, est un indicateur de rentabilité qui correspond à la somme actualisée des profits et pertes futurs générés par les contrats du portefeuille à la date d'évaluation. Elle peut être écrite mathématiquement sous la forme suivante :

$$PVFP = \sum_i \text{Résultat}_i \times DF(0, i)$$

Avec  $\text{Résultat}_i$  le résultat comptable à la date  $i$  et  $DF(0, i)$  le facteur d'actualisation entre les dates 0 et  $i$  tel que :

$$DF(0, i) = \frac{1}{(1 + r)^i}$$

Et  $r$  le taux d'actualisation annuel.

Dans un contexte de simulation par la méthode de Monte-Carlo, le calcul consiste ainsi à projeter sur un nombre d'années définies préalablement le résultat de l'entreprise. Les projections vont dépendre d'un ensemble de scénarios économiques qui auront été définis par un générateur de scénarios économiques (GSE). Les flux seront par la suite actualisés et leur somme permettra d'obtenir la PVFP pour une simulation. Il faudra ainsi calculer la moyenne empirique des PVFP issues des multiples simulations afin d'obtenir la PVFP avec notre modèle.

La PVFP obtenue est un indicateur qui va fortement dépendre des hypothèses de notre GSE, il est donc important d'introduire le fonctionnement de ce générateur.

### GSE

Un générateur de scénarios économiques est un outil qui va permettre de générer des projections de certaines valeurs comme les actions, l'immobilier ou les taux d'intérêts. Ces projections seront effectuées sur un nombre d'années prédéfini. Les scénarios économiques permettront, dans la suite des travaux, d'alimenter un modèle ALM.

Il existe 2 univers de projection utilisables pour un GSE : l'univers risque neutre (RN) et l'univers monde réel (RR). Utiliser un GSE en univers risque neutre implique notamment que le rendement des actifs sera en moyenne égal au rendement de l'actif sans risque. De plus, dans cet environnement les prix sont *market consistent*, soient cohérents avec ceux observés sur le marché. Utiliser les projections d'un GSE RN est nécessaire pour le calcul d'indicateurs sous Solvabilité II, néanmoins l'hypothèse de rendement des actifs qui correspond au rendement de l'actif sans risque, bien que prudente, est éloignée de la réalité des rendements historiques.

Le GSE RR permet quant à lui d'effectuer des projections en ayant pour hypothèses les cours historiques. Le S&P500 dont le rendement annualisé et dividendes réinvesti est de 10,5% depuis 1988 peut par exemple être cité. Bien que depuis les années 2000 ce rendement a diminué, il est possible de voir l'impact d'une telle hypothèse de rendement par rapport au taux sans risque pour un GSE RN. L'univers risque réel permet ainsi d'avoir des produits financiers plus rémunérateurs par l'application d'une prime de risque sur le rendement.

Afin d'utiliser la PVFP comme indicateur du rendement du portefeuille considéré, un GSE RR sera donc utilisé afin d'obtenir des résultats plus proches de ceux espérés par un assureur sur le marché.

Nous pouvons enfin écrire  $PVFP = VM - BEL$ , avec  $VM$  la valeur de marché des actifs du portefeuille et  $BEL$  le *Best Estimate*. Ainsi cette équation peut être traduite par le gain (ou la perte si la PVFP est négative) qui est égal à la différence entre la valeur de marché des actifs et les engagements de l'assureur. Dans notre contexte, nous pouvons comprendre l'intérêt de la PVFP sous univers risque réel car les projections de la VM de nos actifs seront en majeure partie attribuées aux cotisants, et le reste sera le fruit du travail de l'assureur. Utiliser ce critère à maximiser afin d'en extraire la performance de l'assureur lors de la projection de son bilan permettra d'avoir un indicateur de rentabilité cohérent.

L'indicateur de rendement dans le cadre de cette étude a été choisi. Il faut désormais sélectionner un indicateur de risque qui permettra d'obtenir un couple rendement-risque pertinent. Le SCR de marché semble être un candidat idéal afin de remplir ce rôle.

## 1.4 Le SCR de marché

### 1.4.1 Contexte général de la réglementation Solvabilité II

La particularité des assurances d'avoir un cycle de production inversé leur impose une gestion prudentielle de leurs activités. Dans ce but, la directive Solvabilité I a été mise en place en 2002. Assez rapidement, bon nombre de limites ont été mises en avant. Les différents actifs étant évalués avec une vision comptable, les corrélations entre ces actifs n'étaient pas non plus prises en compte, cette approche prudentielle ne permettait pas de s'assurer de la solvabilité des assurances en cas d'évènements extrêmes. Afin de palier à cela, Solvabilité II est venu remplacer Solvabilité I à partir de 2016.

La réforme Solvabilité II implique de grands changements dans le secteur de l'assurance. Ayant pour objectif principal de prémunir toutes les entreprises des risques de faillite (tout comme Solvabilité I), un accent est mis sur la possibilité de contrôle par les autorités au niveau Européen, afin d'identifier les cas problématiques des entreprises présentes sur le marché et donc de mieux protéger les assurés. C'est une approche basée sur le risque qui doit permettre aux entreprises d'être en mesure d'absorber les chocs auxquels elles peuvent être confrontées. La mesure renforce surtout les exigences en terme de fonds propres et de gestion des risques.

La réforme est structurée autour de 3 piliers.

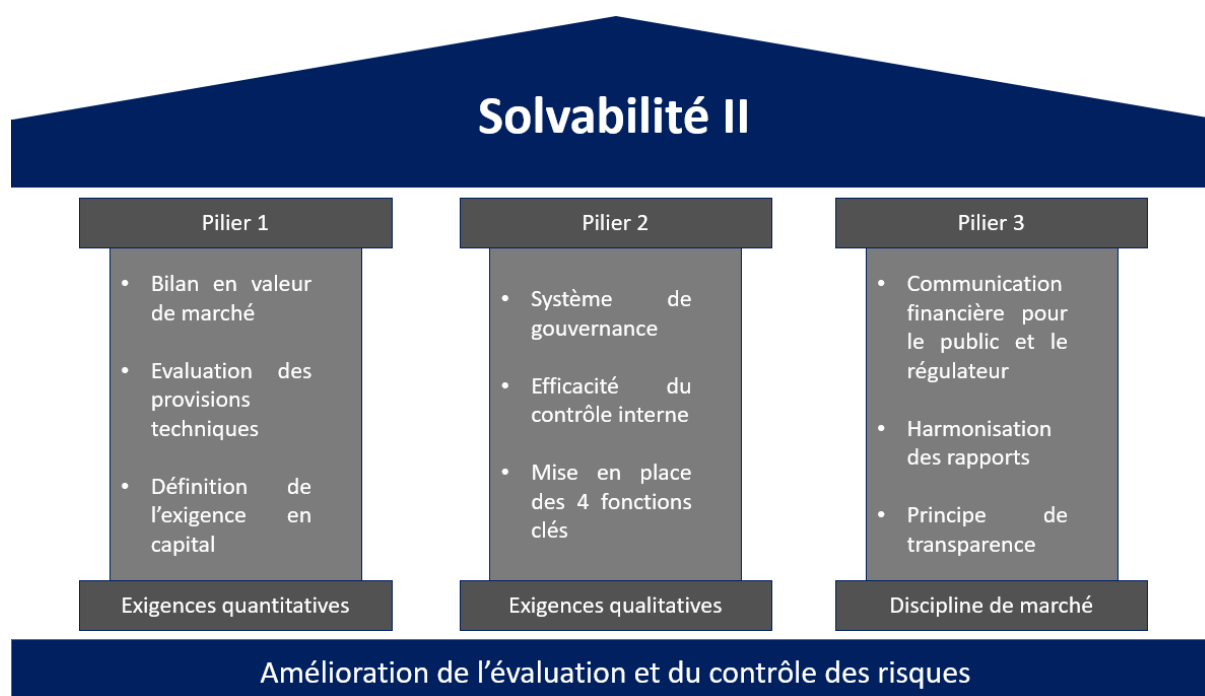


Figure 4 - Piliers de la réforme Solvabilité II

#### Pilier 1

Le pilier 1 concerne les normes quantitatives fixées par l'EIOPA sur les provisions techniques et les fonds propres. Deux indicateurs réglementaires ont ainsi été mis en place :

- Le SCR, *Solvency Capital Requirement*, qui représente le capital nécessaire afin d'absorber les chocs provoqués par les différents risques.

- Le MCR, *Minimum Capital Requirement*, qui représente le niveau minimum de fonds propres en dessous duquel l'autorité de contrôle devra intervenir afin de définir quand et comment l'entreprise pourra à nouveau respecter le seuil de solvabilité réglementaire.

De plus, Solvabilité II impose une évaluation des actifs et passifs en valeur de marché, la réglementation ne repose plus sur un bilan comptable mais économique avec la prise en compte de l'état du marché. Les provisions techniques sont ainsi définies comme la somme du BE, *Best Estimate*, c'est-à-dire la meilleure estimation des flux futurs, et de la RM, *Risk Margin*, soit la marge de risque qui prend en compte les incertitudes du BE.

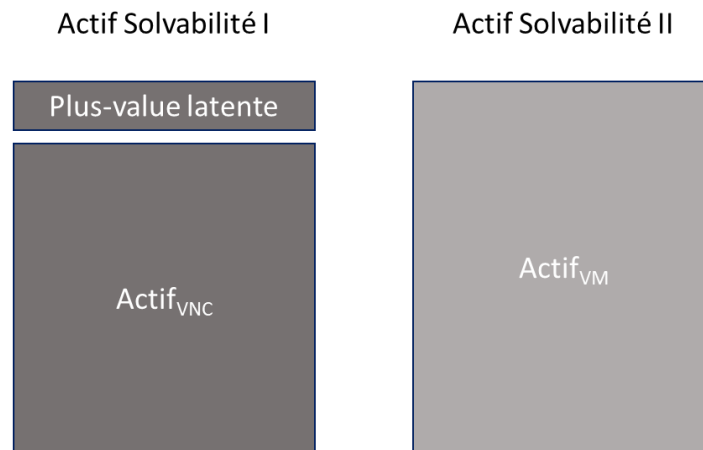


Figure 5 - Passage de l'actif sous Solvabilité II

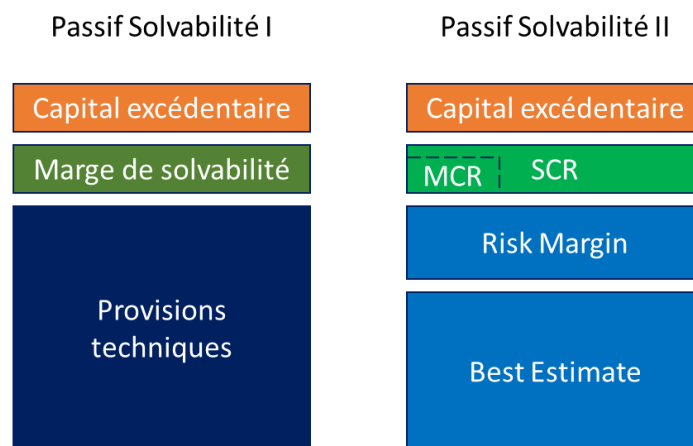


Figure 6 - Passage du passif sous Solvabilité II

Afin d'obtenir notre SCR, il faut calculer notre *Value at Risk* (VaR) qui correspond au montant des pertes entraînant la faillite avec une probabilité définie à 0,5% d'après la réglementation, donc avoir un risque de faillite pour l'entreprise d'une fois tous les 200 ans au maximum.

Le SCR doit prendre en compte plusieurs modules de risques définis selon le profil de l'entreprise en modèle de calcul interne, ou selon la directive en formule standard. La formule standard recense la liste des risques susceptibles d'impacter une compagnie d'assurance.

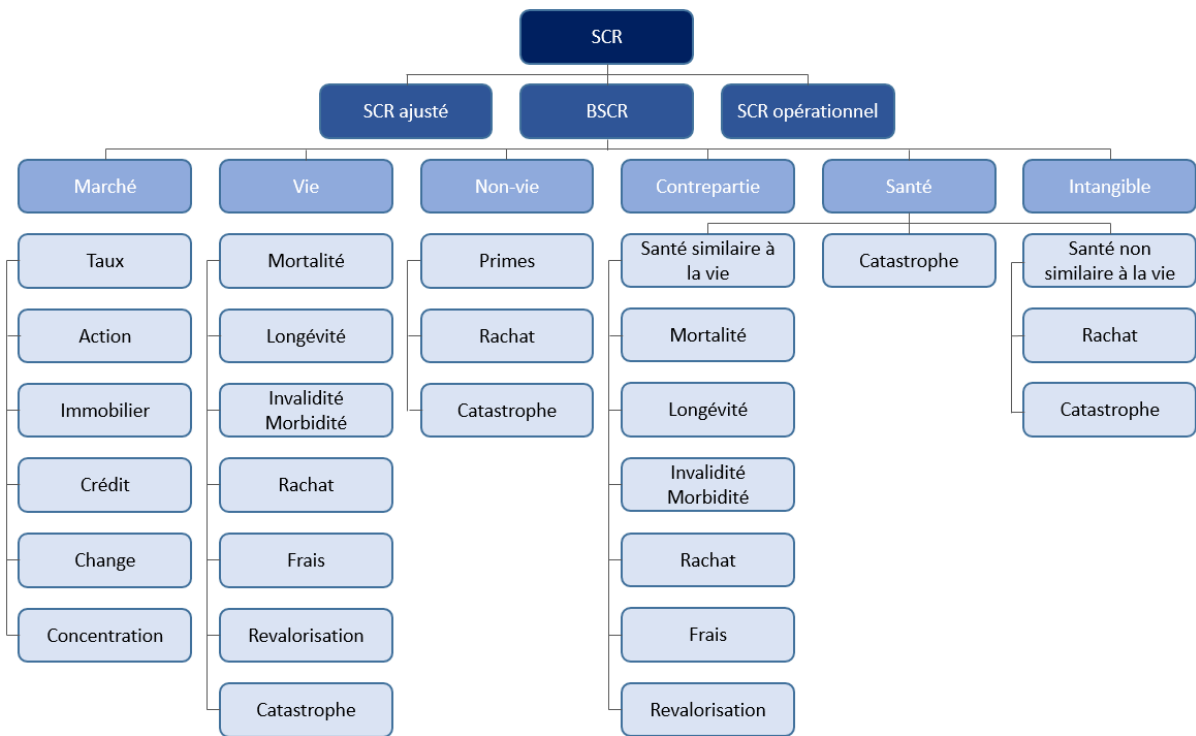


Figure 7 - Pieuvre du SCR

La cartographie des risques ci-dessus communiquée par l’EIOPA permet de représenter les différents modules de risque comme le regroupement de différents sous-modules. En général, le SCR des différents sous-modules est calculé en appliquant un stress sur le scénario central afin de mesurer l’impact du choc sur le BE. Dans la formule standard, le SCR associé à un risque correspond à l’écart entre le bilan en scénario central et le bilan en scénario choqué.

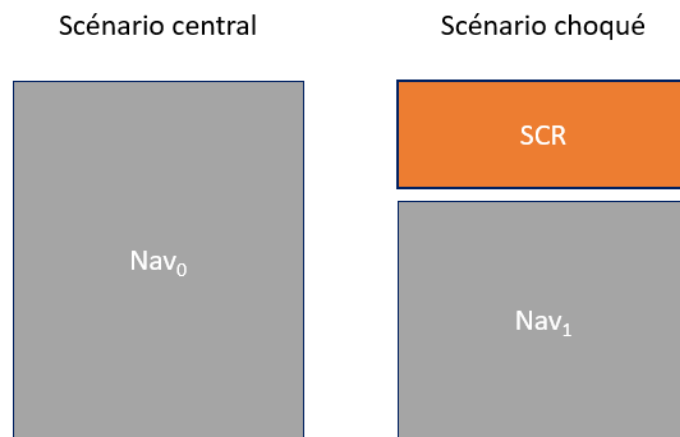


Figure 8 - Représentation du SCR en scénario choqué

La NAV (*Net Asset Value*) correspond à la valeur de l’actif net, et se retrouve ainsi diminuée suite à l’application du choc.

Pilier 2

Le pilier 2 concerne la gouvernance et la surveillance du système de gestion des risques. Ainsi, 4 fonctions clés ont été définies avec des règles très précises concernant la transparence afin de s’assurer de la gestion saine et prudente de l’assurance.

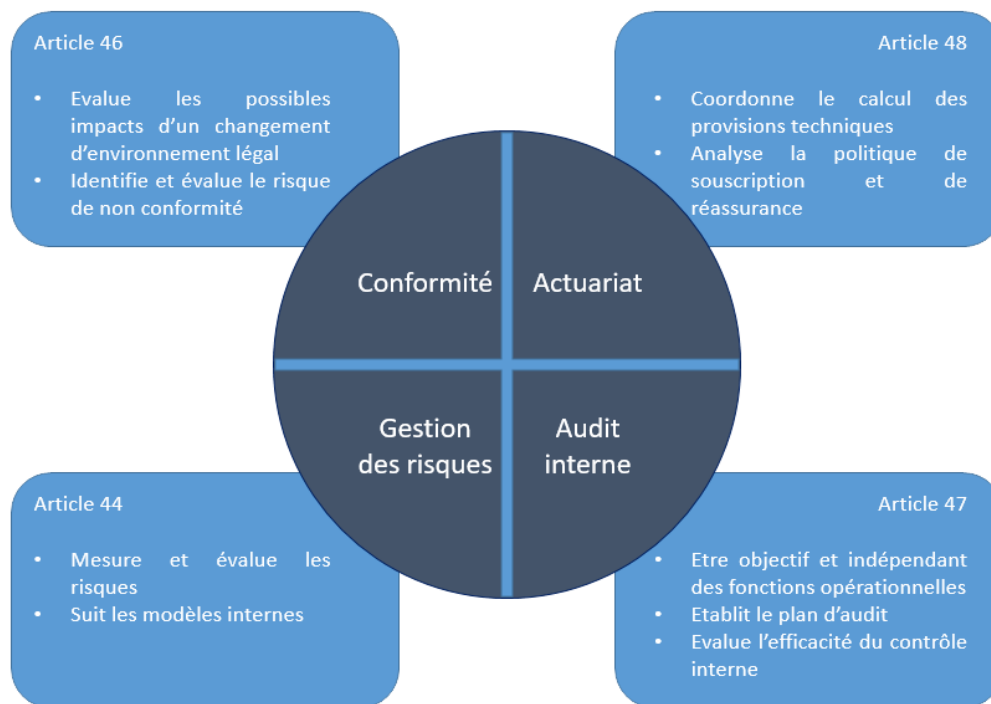


Figure 9 - Fonctions clés du pilier 2 de Solvabilité II

- La fonction de gestion des risques s'occupe des mesures à mettre en place afin de gérer les risques assimilés à l'entreprise.
- La fonction conformité qui veille au respect des règles légales et administratives de Solvabilité II.
- La fonction actuarielle qui vérifie la fiabilité des calculs liés aux provisions techniques.
- La fonction audit interne qui est indépendante des autres et a pour objectif de s'assurer de l'efficacité du système de contrôle interne.

De nombreuses règles sont en place afin de s'assurer de la qualité des fonctions clés. Tout d'abord, par rapport au choix des responsables de chaque fonction, ces derniers doivent attester d'un certain niveau de compétence. Cela peut être effectué par la vérification des diplômes que détient le membre, mais aussi de ses connaissances, qu'elles correspondent à ce qui est attendu dans la fonction. L'honorabilité du responsable est aussi un point important d'autant plus en ce qui concerne la fonction d'audit interne qui se veut indépendante du reste de l'entreprise. Le cumul des responsabilités est à éviter, c'est-à-dire que les membres des fonctions clés ne peuvent pas avoir de rôle dans l'entreprise impactant leur objectivité. Par exemple un membre de la fonction audit interne ne devra pas faire d'audit sur des travaux sur lesquels il aurait participé dans la période auditée.

Les règles sont néanmoins adaptables en fonction du contexte de l'entreprise. Le principe de proportionnalité indique que les entreprises peuvent mettre en place les directives en accord avec leurs besoins et contraintes. L'objectif étant d'une part, de permettre une gestion des risques qualitative, mais tout en évitant de mettre en danger l'activité économique de l'entreprise ou son organisation. Certaines entreprises pourront ainsi confier par exemple les fonctions de conformité et de gestion des risques à la même équipe afin de réduire les coûts en ressource humaine et d'éviter les coûts disproportionnés par rapport à la taille de l'entreprise. Tout fonctionnement de façon simplifié devra néanmoins être justifié auprès du régulateur.

### Pilier 3

Ce pilier concerne la transparence des risques, c'est-à-dire l'information publique qui doit renforcer la discipline de marché. L'objectif est ainsi d'harmoniser les informations transmises dans les Etats membres au niveau des assurés, du marché et des autorités de contrôle.

L'ensemble des assureurs européens doivent transmettre des rapports au format plus ou moins similaire mais avec une fréquence et un contenu harmonisé afin de permettre une comparaison simplifiée entre les acteurs. Ces rapports devront contenir les informations nécessaires et vérifiables afin de s'assurer de la santé financière des compagnies ou afin de les superviser.

Plusieurs types de documents doivent ainsi être transmis par les compagnies :

- QRT, *Quantitative Reporting Template*, soient les états quantitatifs, tous les trimestres avec une version consolidée tous les ans qui doit être transmise au régulateur.
- SFCR, *Solvency and Financial Conditions Reporting*, au même rythme, et qui correspondent à des rapports narratifs sur la solvabilité et la situation financière à destination du public et du régulateur.
- RSR, *Regular Supervisory Reporting*, qui correspond aussi à des rapports narratifs réguliers au contrôleur, dans les mêmes conditions que le SFCR.

Le cadre de ce mémoire se situe dans le pilier 1 de la directive. Nous nous intéresserons au SCR comme mesure de risque et plus particulièrement au SCR de marché dans un contexte d'allocation d'actifs.

#### 1.4.2 Calcul du SCR de marché

L'article 105 (5) de la directive 2009/138/CE du Parlement Européen définit le module de risque de marché de la façon suivante :

Le module « risque de marché » reflète le risque lié au niveau ou à la volatilité de la valeur de marché des instruments financiers ayant un impact sur la valeur des actifs et des passifs de l'entreprise concernée. Il reflète de manière adéquate toute inadéquation structurelle entre les actifs et les passifs, en particulier au regard de leur durée.

Ainsi, il est composé des sous-modules suivants :

- Taux, *Interest rate* : risque de baisse ou hausse des taux.
- Action, *Equity* : risque de chute instantanée du cours des actions.
- Immobilier, *Property* : risque de chute instantanée des prix de l'immobilier.
- *Spread* : notamment le risque de baisse de qualité de crédit de l'émetteur.
- Change, *Currency* : risque de baisse des taux de change par rapport à la devise de référence.
- Concentration : risque de non-diversification des actifs.

Le calcul du SCR de marché (*Market*) fait intervenir les SCR propre à chaque sous-module en prenant en compte leur coefficient de corrélation. Nous avons ainsi la formule suivante :

$$SCR_{Market} = \sqrt{\sum_{i,j} Corr_{i,j} \cdot SCR_i \cdot SCR_j}$$

Avec  $SCR_i$  et  $SCR_j$  les sous-modules de marché et  $Corr_{i,j}$  leur coefficient de corrélation. La matrice de corrélation suivante permet d'attribuer au couple de sous-modules du SCR le coefficient défini par la réglementation :

	Interest rate	Equity	Property	Spread	Currency	Concentration
Interest rate	100%	0% 50%	0% 50%	0% 50%	25%	0%
Equity	0% 50%	100%	75%	75%	25%	0%
Property	0% 50%	75%	100%	50%	25%	0%
Spread	0% 50%	75%	50%	100%	25%	0%
Currency	25%	25%	25%	25%	100%	0%
Concentration	0%	0%	0%	0%	0%	100%

	En cas de risque des taux à la hausse
	En cas de risque des taux à la baisse

Figure 10 - Matrice de corrélation du SCR de marché

### Module risque de taux

Ce module vise à quantifier le besoin en capital nécessaire pour faire face à l'impact d'une évolution de la structure de la courbe des taux (à la hausse ou la baisse) sur la valeur du bilan. Le choc de taux est en l'occurrence égal au maximum des chocs à la hausse et à la baisse de la courbe des taux.

Les principaux actifs concernés par ce module sont les obligations d'entreprises et d'état, mais aussi les produits dérivés tels que les options, futures ou swaps. La courbe des taux permet de définir les chocs à la hausse et à la baisse en fonction de la maturité. Le choc est un pourcentage du niveau des taux, et ne peut être inférieur à 1% dans le cas de hausse des taux, ou est nul lors d'un choc à la baisse si les taux sont négatifs.

Plusieurs approches existent afin de calculer le SCR de taux. L'approche par revalorisation est la plus utilisée, elle consiste à faire la différence entre un prix théorique obtenu après application du choc et le prix en scénario central. C'est une très bonne approche en général qui est aussi adaptée aux produits ayant un *pay-off*, c'est-à-dire un résultat, complexe ou intégrant des optionalités. Il faut néanmoins avoir à disposition toutes les caractéristiques des actifs ainsi que des données de marchés pour donner un prix à chaque produit.

L'approche par la sensibilité est une seconde façon de calculer le SCR de taux. Cela consiste à projeter le choc sur l'exposition ajustée de la sensibilité au risque de taux (*Modified Duration* noté *ModDur*). L'avantage de cette méthode est qu'elle nécessite moins de données, néanmoins sur un instrument fortement convexe, cette approche va diverger de celle par



revalorisation. Par exemple, pour les obligations vanilles, cela va majorer le SCR de taux à la hausse et minorer le SCR de taux à la baisse. On a ainsi la formule suivante :

$$SCR_{Taux} = Exposition \times (\Delta r \times ModDur)$$

Avec  $\Delta r$  le niveau de choc.

La dernière approche est similaire à la précédente en prenant en compte la convexité. Il faut donc introduire la convexité de taux  $Cvx$  afin de réduire l'écart de divergence. Cette méthode est notamment nécessaire pour les chocs de forte amplitude et les instruments ayant une sensibilité de taux importante. On a désormais :

$$SCR_{Taux} = Exposition \times (\Delta r \times ModDur - \Delta r^2 \times \frac{Cvx}{2})$$

### Module risque action

Ce module vise à quantifier l'impact d'une baisse soudaine des marchés actions (risque action) sur la valeur du bilan de l'assureur. Il faut noter qu'il prend aussi en compte les titres directement impactés par l'évolution des marchés actions (futures sur indice action, options sur action ou indice action, warrants sur action, obligations convertibles).

Il est décomposé en 2 catégories :

- Les actions de type 1 (Type1equities), soit les actions listées sur un marché réglementé ou échangées sur une plateforme multilatérale de négociation dans un pays membre de l'UE ou de l'OCDE
- Les actions de type 2 (Type2equities) qui correspond à toutes les actions non comprises dans le type 1.

Le SCR se calcule comme une perte induite par le choc instantané à la baisse applicable à chacune des catégories :

$$\begin{aligned} \text{Type1equities} &= 39\% + SA \\ \text{Type2equities} &= 49\% + SA \end{aligned}$$

Avec  $SA$  qui correspond au *Symmetric Adjustment*. C'est un facteur d'ajustement contra-cyclique qui évolue entre -10% et +10%. Il est publié mensuellement par l'EIOPA.

$$SA = \frac{1}{2} \left( \frac{CI - AI}{AI} - 8\% \right)$$

Avec  $CI$  le niveau actuel de l'indice et  $AI$  la moyenne journalière sur 3 ans de l'indice. Enfin, le SCR peut être calculé de la façon suivante :

$$SCR_{Action} = \sqrt{0,75 \times 2 \times SCR_{type1equities} \times SCR_{type2equities} + SCR_{type1equities}^2 + SCR_{type2equities}^2}$$

La méthodologie de calcul du SCR action se rapproche de celles évoquées pour le module de risque de taux.

### Module risque immobilier

Ce module vise à quantifier l'impact de la baisse des marchés immobiliers sur la valeur des actifs et consiste à faire baisser instantanément la valeur des actifs immobiliers de 25%. Les terrains, constructions ou droit sur des biens immobiliers ainsi que les investissements immobiliers détenus pour utilisation par une entreprise sont assimilés au risque immobilier. Les sociétés ayant une activité de gestion d'installations, administration d'immeubles, conception de projets immobiliers ou activités similaires sont quant à eux assimilés au risque action.

En fonction de la classe d'actif le choc appliqué peut varier légèrement. En effet pour l'immobilier en direct nous avons :

$$SCR_{Immobilier} = Exposition \times 75\%$$

Les SCPI, SCI, OPCI sont calculés différemment :

$$SCR_{Immobilier} = Exposition \times 75\% \times \text{taux\_investissement}$$

### Module risque de spread

Ce module vise à quantifier le besoin en capital correspondant à une hausse (ou une baisse pour le CDS, *Credit Default Swap*) de l'écart entre le taux actuariel d'un produit de taux et le taux sans risque de la devise du titre. Il est divisé en 3 catégories et on a :

$$SCR_{spread} = SCR_{bonds} + SCR_{securisation} + SCR_{cd}$$

Avec :

- $SCR_{bonds}$  concerne les obligations, les prêts, les comptes à termes et dépôts à termes.
- $SCR_{securisation}$  concerne les titres issus de titrisation.
- $SCR_{cd}$  concerne les dérivés de crédit CDS et CLN ou les TRS sur indice crédit.

Il faut donc calculer de façon indépendante les 3 SCR qui composent le  $SCR_{spread}$ .

$$SCR_{bonds} = \sum_i VM_i \cdot Stress_i(\text{credit quality step}_i, \text{duration}_i)$$

Avec :

- $VM_i$  la valeur de marché de l'obligation
- $duration_i$  la sensibilité de crédit de l'obligation (toujours  $\geq 1$ )
- $credit\ quality\ step_i$  la notation du crédit
- $Stress_i$  la fonction de stress qui est calculée à l'aide de la matrice suivante :

		Credit quality step						
		0	1	2	3	4	5/6	NR
Duration	0 ~ 5	0.9% × duration	1.1% × duration	1.4% × duration	2.5% × duration	4.5% × duration	7.5% × duration	3% × duration
	5 ~ 10	4.5% + 0.5% × (duration - 5)	5.5% + 0.6% × (duration - 5)	7% + 0.7% × (duration - 5)	12.5% + 1.5% × (duration - 5)	22.5% + 2.5% × (duration - 5)	37.5% + 4.2% × (duration - 5)	15% + 1.7% × (duration - 5)
	10 ~ 15	7% + 0.5% × (duration - 10)	8.5% + 0.5% × (duration - 10)	10.5% + 0.5% × (duration - 10)	20% + 1% × (duration - 10)	35% + 1.8% × (duration - 10)	58.5% + 0.5% × (duration - 10)	23.5% + 1.2% × (duration - 10)
	15 ~ 20	9.5% + 0.5% × (duration - 15)	11% + 0.5% × (duration - 15)	13% + 0.5% × (duration - 15)	25% + 1% × (duration - 15)	44% + 0.5% × (duration - 15)	61% + 0.5% × (duration - 15)	23.5% + 1.2% × (duration - 10)
	Supérieur à 20	12% + 0.5% × (duration - 20)	13.5% + 0.5% × (duration - 20)	15.5% + 0.5% × (duration - 20)	30% + 0.5% × (duration - 20)	46.5% + 0.5% × (duration - 20)	63.5% + 0.5% × (duration - 20)	Min(35.5% + 0.5% × (duration - 20) ; 1)

Figure 11 - Matrice de choc de spread pour le SCR bonds

Il est à noter que dans certains cas, la matrice utilisée afin de déterminer le niveau de choc peut être différente. C'est le cas notamment des obligations émises par des Etats non-membres de l'EEA en devise locale et des obligations garanties (*covered bonds*). L'étape suivante consiste à calculer le  $SCR_{securisation}$  grâce à la formule :

$$SCR_{securisation} = \sum_i VM_i \cdot Stress_i(credit\ quality\ step_i, duration_i, type_i)$$

Avec :

- $type_i$  = Type 1 pour les actifs ayant une notation BBB- ou plus (échelon 3), listés sur un marché organisé liquide d'un pays membre de l'EEA ou OCDE, et enfin être issu de la tranche la plus sénior du véhicule de titrisation.
- $type_i$  = Type 2 pour les autres actifs.
- $type_i$  = « Resecurisation » pour les titrisations ayant à l'actif d'autres titrisations.

		Credit quality step					
		0	1	2	3	4	5/6/NR
Duration	Type 1	Min(2.1% × duration ; 1)	Min(3% × duration ; 1)	Min(3% × duration ; 1)	Min(3% × duration ; 1)		
	Type 2	Min(12.5% × duration ; 1)	Min(13.4% × duration ; 1)	Min(16.6% × duration ; 1)	Min(19.7% × duration ; 1)	Min(82% × duration ; 1)	100%
	Resecurisation	Min(33% × duration ; 1)	Min(40% × duration ; 1)	Min(51% × duration ; 1)	Min(91% × duration ; 1)	100%	100%

Figure 12 - Matrice de choc de spread pour le SCR securisation

Pour le calcul du  $SCR_{cd}$ , la formule suivante est utilisée :

$$SCR_{cd} = \max \left\{ \sum_i Choc_{spread\ up_i}, \sum_i Choc_{spread\ down_i} \right\}$$

Avec le niveau de choc qui dépend de la qualité de crédit comme suivant :

Credit quality step	Stress spread up	Stress spread down
0	+130 bp	-75%
1	+150 bp	-75%
2	+260 bp	-75%
3	+450 bp	-75%
4	+840 bp	-75%
5	+1620 bp	-75%
6	+1620 bp	-75%
NR	+500 bp	-75%

Figure 13 - Matrice de choc de spread pour le SCR cd

Dans le cas des CDS, le calcul peut ainsi se faire en suivant une approche par revalorisation, ou bien avec une approche par sensibilité. Dans ce dernier cas, le choc est projeté sur l'exposition ajustée de la sensibilité au risque de crédit (duration). Ainsi nous avons :

$$Choc_{spread\ up} = Stress_{up}(bp) \times Sensibilité\ de\ crédit_{CDS} \times Exposition_{CDS}$$

$$Choc_{spread\ down} = Stress_{down}(\%) \times Spread_{CDS} \times Sensibilité\ de\ crédit_{CDS} \times Exposition_{CDS}$$

Avec :

$$Spread_{CDS} = Coupon_{CDS} + \frac{Valeur\ boursière\ coupons\ exclus_{CDS}}{Nominal_{CDS} \times Sensibilité\ de\ crédit_{CDS}}$$

### Module risque de change

Ce module vise à quantifier le besoin en capital correspondant à la perte générée par l'effet de change sur la valeur des actifs. Cela concerne les titres libellés en devise étrangère, les produits de change (options, futures, swap, forwards) et les produits structurés indexés sur le change. Le stress de change est de 25% sauf pour les devises ayant un ancrage de change avec l'euro. L'exemple peut être pris de la Couronne Danoise pour laquelle le stress de change est d'environ 0.39%, ou alors le Franc Comorien pour lequel le stress de change est d'environ 2%. Le SCR se calcule ainsi de la façon suivante :

$$SCR_{Change} = \sum_{i\ devise} (Stress_{i\ devise} \times |Exposition_{i\ devise}|)$$

### Module risque de concentration

Ce module vise à quantifier le besoin en capital correspondant à un manque de diversification des actifs ou à une surexposition au risque de défaut d'un émetteur (on le calcule au niveau du groupe d'un émetteur). Cela concerne les actifs présents dans les modules risque action, risque immobilier et risque de crédit.

On écrit le niveau de surexposition  $XS_i$  tel que :

$$XS_i = \max \left( 0 ; \frac{E_i}{Assets_{xl}} - CT_i \right)$$

Avec :

- $E_i$  l'exposition nette de risque de défaut d'une entité (et égale à la somme des expositions nette individuelles de cette entité)
- $Assets_{xl}$  la somme des VM des actifs concernés
- $CT_i$  le seuil de surexposition

Cela permet d'obtenir le coût individuel de l'exposition d'un émetteur  $Con_i = XS_i \cdot g_i$  avec  $g_i$  un facteur pénalisant de surexposition. La formule du SCR de concentration est ainsi la suivante :

$$SCR_{conc} = \sqrt{\sum (Con_i)^2}$$

# Approche du problème

## 2.1 Mise en avant de solutions d'allocation d'actifs

Différentes solutions permettant de définir une allocation d'actifs optimale peuvent être utilisées. Parmi les algorithmes permettant de résoudre un problème d'optimisation, la méthode de la descente du gradient ainsi que l'algorithme génétique peuvent être mentionnés.

### 2.1.1 Méthode de descente du gradient

La problématique du choix de l'allocation d'actif optimal peut être associée à un problème d'optimisation comme nous avons pu le voir précédemment. Afin de résoudre ce problème, la méthode de descente du gradient peut être appliquée.

Cette méthode consiste en l'application d'un algorithme afin de minimiser ou maximiser une fonction cible. Cette optimisation est le résultat d'améliorations successives, soit le produit d'un grand nombre d'itérations. A chacune de ces itérations, nous passons d'une allocation à une autre en se rapprochant de l'optimum. Il est possible de représenter une telle fonction sur un espace en 3 dimensions comme ci-dessous :

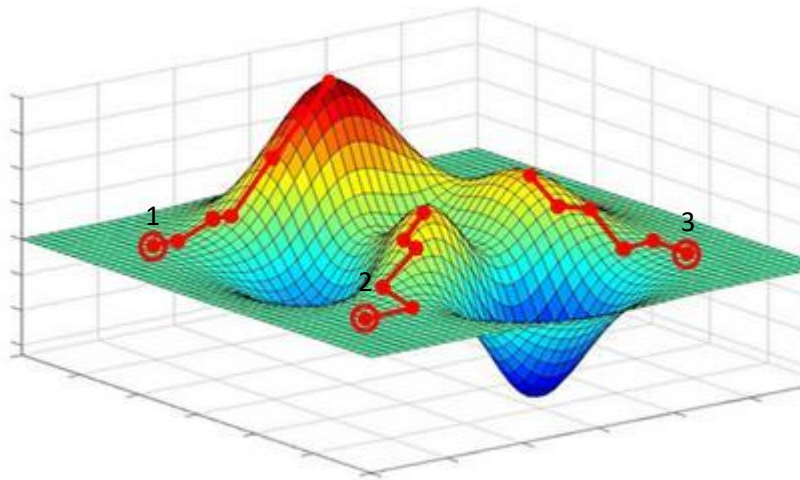


Figure 14 - Représentation d'un plan de solution sur 3 dimensions issu du site : <https://jcrisch.wordpress.com/2015/04/02/les-reseaux-de-neurones/>

Ce plan correspond à l'ensemble des valeurs possible pour notre fonction d'optimisation. Les points rouges encerclés correspondent à l'allocation de départ de l'algorithme. Si l'objectif est de maximiser notre fonction, on peut observer le comportement des 3 trajectoires dont le résultat final va fortement dépendre de la fonction et de l'allocation initiale. Les trajectoires 2 et 3 trouvent rapidement un maximum local néanmoins la maximisation sera plus efficace en suivant la trajectoire 1.

L'enjeu de cet algorithme sera donc de réussir à atteindre l'optimum global de notre fonction sans être piégé par un optimum local au fil des itérations. Le choix du pas d'itération est ainsi primordial. La direction du déplacement est quant à elle obtenue avec le calcul du gradient qui indique la direction de la pente la plus forte à partir d'un point donné.

Si nous prenons l'exemple d'une fonction de maximisation de la PVFP ayant comme variable l'allocation des différentes classes d'actifs, nous pourrions utiliser l'algorithme suivant :

1. Initialisation de l'allocation initiale et calcul de la PVFP
2. Modification de l'allocation initiale d'un actif par rapport à un pas que l'on définit (par exemple une augmentation du taux d'action de 1% contre une baisse de la somme des taux des autres actifs de 1%)
3. Calcul de la PVFP
4. Sélection de l'allocation maximisant la PVFP, et répétition au point numéro 2

Cette méthode comprend néanmoins certaines limites qui la rendent moins évidente à utiliser en pratique. Tout d'abord, c'est un modèle d'optimisation sans contrainte ce qui peut mener à des allocations incohérentes et injustifiables en pratique. Il est notamment possible d'obtenir des allocations négatives sur certains actifs ou excessives sur d'autres.

Si on reprend l'exemple précédent dont la fonction chercherait à maximiser la PVFP, sans contrainte nous pourrions avoir le taux d'allocation d'action qui serait beaucoup trop grand par rapport aux autres actifs. Cela pourrait aussi amener un manque de diversification. Afin de palier à cela et dans le but de l'optimisation du couple rendement-risque d'un portefeuille, il pourrait être judicieux d'intégrer des contraintes pénalisantes en fonction du risque associé à une allocation testée.

Malgré l'intégration de contraintes dans le problème d'optimisation, il reste toujours le risque d'être piégé par un optimum local. Cela peut néanmoins être évité en partie en utilisant l'algorithme plusieurs fois avec différentes allocation initiales.

Enfin, chaque itération impliquant un nouveau calcul de la fonction objectif, le temps de calcul peut vite devenir très long, notamment si le critère d'optimisation prend en compte des indicateurs de Solvabilité II comme la PVFP ou le SCR.

## 2.1.2 Algorithme génétique

Une approche alternative pouvant être mise en place dans le cadre de résolution d'un problème d'optimisation, pour la détermination d'une allocation d'actif optimale, est l'utilisation d'un algorithme génétique. Basé sur un phénomène biologique, il s'inspire de la théorie de l'évolution qui décrit qu'au fil du temps, les gènes les plus adaptés à l'environnement d'un individu seront conservés au sein d'une population. L'algorithme cherche à reproduire cette théorie, c'est-à-dire garder l'information des meilleurs individus en les améliorant au fil des générations, ou itérations.

Nous avons donc une population qui consiste en un ensemble de solutions, dont l'évolution peut être simulée afin d'obtenir les meilleurs individus ou solutions. Afin de modéliser cette évolution, trois opérateurs interviennent et sont illustrés pour des individus en représentation binaire :

- La sélection, qui consiste à déterminer les individus retenus et à partir desquels de nouvelles solutions vont être générées. Il existe plusieurs méthodes de sélection, que ce soit entièrement aléatoires ou bien en fonction du rang des individus, soit ceux qui sont les plus performants par rapport à notre fonction objectif.

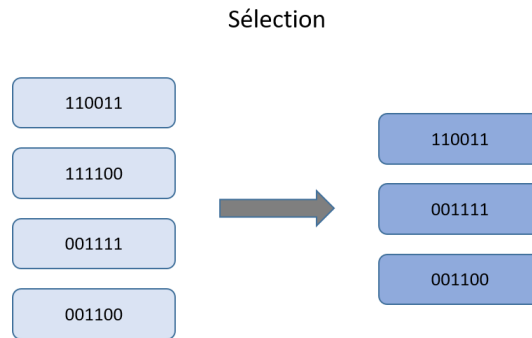


Figure 15 - Illustration de l'opérateur de sélection

- Le croisement dont l'objectif est d'imiter le principe d'hérédité de la théorie de Darwin. Certaines parties des individus sélectionnés précédemment sont mélangées afin de créer des nouveaux individus.

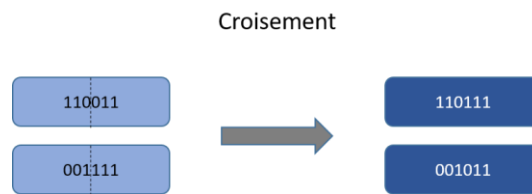


Figure 16 - Illustration de l'opérateur de croisement

- La mutation est le dernier opérateur qui consiste à faire varier un élément d'un individu et permet en l'occurrence d'éviter une situation de blocage dans un optimum local.



Figure 17 - Illustration de l'opérateur de mutation

Dans le cas d'un algorithme cherchant à déterminer une allocation optimale d'actif en se basant sur les critères PVFP et SCR, les bits des individus illustrant les opérateurs ci-dessus pourraient correspondre à l'allocation choisie pour chaque actif. La performance de chaque individu pourrait éventuellement être la maximisation du ratio PVFP/SCR sous contrainte d'un SCR maximal. L'algorithme pourra finalement se dérouler de la façon suivante :

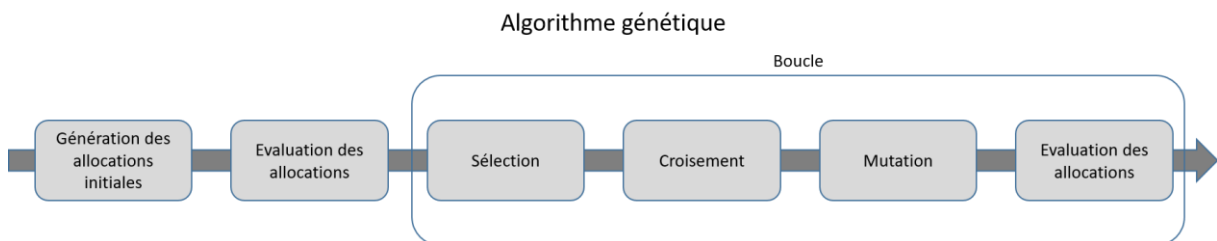


Figure 18 - Démarche d'un algorithme génétique

La boucle pourra être effectuée jusqu'à ce qu'un critère d'arrêt soit atteint.

Finalement, l'algorithme génétique présente des avantages dans le cadre d'une problématique d'allocation d'actifs, notamment car il permet de réduire le risque d'être bloqué dans un optimum local. De plus, la convergence vers l'optimum est en général rapide.



Néanmoins bien que performant par rapport à d'autres méthodes d'optimisation, l'algorithme génétique implique une complexité de calcul très importante. Le calcul de la PVFP et du SCR étant longs pour un modèle ALM, l'algorithme génétique sera long lors de l'étape de l'évaluation des allocations.

De plus, son efficacité dépendra grandement de la chance d'avoir un individu performant dès la génération des allocations initiales.

### 2.1.3 Analyse critique

Bien que les méthodes présentées ci-dessus puissent être efficaces afin de déterminer l'allocation d'actifs optimale d'un portefeuille, le temps de calcul nécessaire à leur utilisation est une contrainte de taille pour les assureurs. La méthode de calcul classique consisterait à tester tout l'univers des allocations pour déterminer celle qui est optimale sur un plan rendement-risque en prenant en compte le SCR. La contrainte principale à cette approche est le temps de calcul nécessaire à sa mise en place.

Afin de répondre à cette difficulté opérationnelle, nous nous posons la question de l'utilisation d'une méthode basée sur l'apprentissage automatique aussi appelé *machine learning* dans ce cadre. En effet, bien que la construction des modèles nécessite une base de données conséquente, leur utilisation pourrait permettre de rapidement obtenir la prédiction d'un indicateur de rentabilité ou de risque en fonction de l'allocation testée en *input*.

Dans le cadre de ce mémoire, nous mettrons en pratique des modèles de *machine learning* afin d'évaluer si cette solution permet de répondre à notre problématique de façon rapide et efficace. Le *machine learning* permettra de déterminer une allocation d'actif optimal avec comme mesure de risque le SCR de marché, et comme mesure de rendement la PVFP en scénario monde réel.

## 2.2 Focus sur le *machine learning*

Le *machine learning* est un champ d'étude de l'intelligence artificielle dont le principe est d'apprendre à partir de données. Les modèles de *machine learning* ont une approche statistique et ont pour objectif de résoudre des tâches sans être explicitement programmés pour cela. Ce sont des algorithmes que l'on appelle « entraîables » car ils sont capables de restituer une règle mathématique qui définit les données en s'entraînant sur des exemples. Concrètement, il s'agit de techniques permettant de découvrir des schémas dans une base de données afin d'effectuer des prédictions.

L'efficacité de ce type de modèle est étroitement liée à la qualité des données sur lesquelles il sera basé. En effet le *machine learning* est particulièrement efficace lorsqu'il est utilisé dans un contexte de *Big Data*. L'utilisation de méthodes statistiques classiques dans des jeux de données comprenant un grand nombre d'individus et de variables n'est pas toujours possible, c'est là que l'exploitation de modèles d'apprentissage automatique prend son sens.

L'utilisation d'une base de données aux dimensions plus réduites n'est néanmoins pas contradictoire à l'utilisation de ces méthodes. Bien qu'il soit plus difficile d'entraîner un modèle sur une base comprenant un nombre limité d'individus, certains algorithmes permettent tout de même d'obtenir des résultats pertinents.

La démarche afin de mener à bien un projet en *machine learning* est en générale plus ou moins similaire quel que soit le modèle utilisé et suit les étapes suivantes.

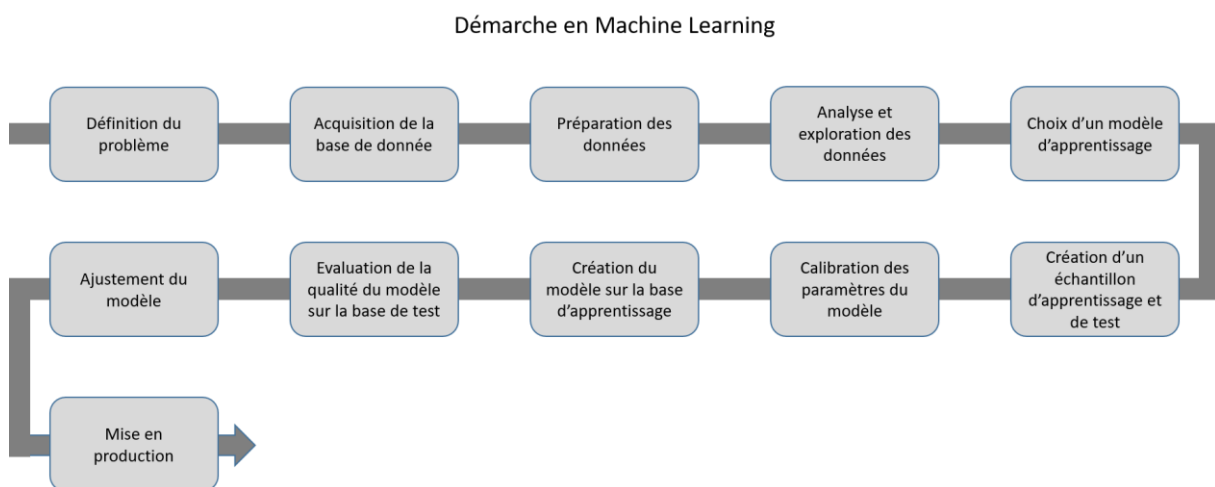


Figure 19 - Démarche d'un projet de *machine learning*

Chaque étape est importante afin de mener à bien un projet de *machine learning*. Le travail préliminaire sur les données va permettre notamment d'identifier le ou les algorithmes qu'il serait pertinent d'utiliser, et de faciliter l'action du modèle. La performance de l'apprentissage automatique étant directement dépendante de la qualité des données en entrée, une importante partie du travail doit être concentrée sur les données.

La calibration des paramètres du modèle est aussi un point déterminant du processus. Chaque algorithme possède un certain nombre d'hyper-paramètres qui vont être utilisés pour contrôler le processus d'apprentissage. Le choix de ces paramètres est souvent réalisé en appliquant une méthode de *cross-validation*.

La phase d'évaluation de la qualité du modèle sur la base de test va permettre de s'assurer de son efficacité sur des données qu'il ne connaît pas. Cela permettra aussi de comparer les différents modèles réalisés en se basant sur différents critères comme l'erreur de prédiction.

L'utilisation de l'apprentissage automatique est de plus en plus grande de nos jours. On peut citer comme exemple les moteurs de recommandations utilisés par des entreprises comme Amazon ou Netflix, mais aussi les moteurs de recherche web comme Google. Le métier de ces entreprises leur permet d'amasser une grande quantité de données qui deviennent exploitables avec des techniques de *machine learning*.

Les voitures autonomes fonctionnent aujourd'hui en utilisant des techniques de machine learning. Bien que les performances soient actuellement limitées, la technologie s'améliore constamment en cherchant à limiter au maximum les erreurs.

Dans le cadre de nos travaux, l'application d'un modèle de *machine learning* visera à effectuer une prédiction des indicateurs PVFP et SCR en fonction de différentes variables telles que l'allocation des actifs. Deux grands types d'apprentissage peuvent ainsi être utilisés, l'apprentissage supervisé et l'apprentissage non-supervisé.

## 2.3 Apprentissage non-supervisé et supervisé

Ces deux types d'apprentissage ont un mode de fonctionnement différent et permettent d'obtenir des résultats différents. Le choix des modèles à appliquer va donc dépendre du problème traité. Il est à noter qu'il existe d'autres catégories comme l'apprentissage par renforcement, néanmoins ils ne seront pas traités dans ce mémoire.

### Apprentissage non-supervisé

L'apprentissage non-supervisé consiste à former des groupes d'individus dans une base de données sans avoir de variable  $Y$  à expliquer lors de l'apprentissage. L'algorithme cherche donc à identifier les individus ayant des caractéristiques similaires et à les regrouper selon un certain nombre de classes qui peut optionnellement être défini au préalable.

Les algorithmes d'apprentissage non-supervisé sont souvent utilisés afin de produire un partitionnement de données, des *clusters*. Les individus sont triés dans les *clusters* en fonction de leur similarité qui est généralement calculée selon une fonction de distance.

#### Clustering avec un apprentissage non-supervisé

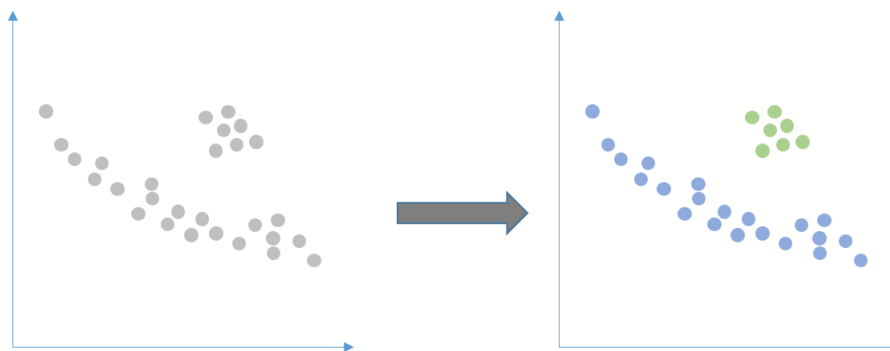


Figure 20 - Classification d'individus lors d'un apprentissage non-supervisé

L'utilisation de l'apprentissage non-supervisé permet de créer des groupes d'individus en fonction de leur distance.

### Apprentissage supervisé

Les données utilisées dans ce type d'apprentissage sont constituées de variables explicatives  $X = \{X_1, X_2, \dots, X_p\}$  ainsi que de la variable à expliquer  $Y$ . Le principe de fonctionnement des algorithmes d'apprentissage supervisé est de s'appuyer sur les individus dont la valeur de  $Y$  est connue afin de construire un modèle permettant d'obtenir  $Y$  à partir de  $X$ .

Nous pouvons distinguer deux classes d'algorithmes d'apprentissage supervisé. Les algorithmes de classification dont l'objectif est de prédire  $Y$  lorsqu'il s'agit d'une variable qualitative. Les algorithmes permettant de réaliser une régression afin de prédire  $Y$  sont utilisés lorsqu'il s'agit d'une variable quantitative.

L'objectif de l'apprentissage supervisé est de créer une fonction qui permet de bien généraliser l'information qui lie  $X$  à  $Y$ , c'est-à-dire qui permet de prédire correctement  $Y$  sur des individus non présents dans la base d'apprentissage en fonction de leurs caractéristiques  $X$ . La notion de généralisation est importante afin d'éviter le phénomène de sur-apprentissage. En effet si le modèle

réalisé est « trop » entraîné sur notre base de données, il peut arriver qu'il permette de prédire  $Y$  avec une erreur minimale, sans pour autant être efficace sur de nouvelles observations. Le sur-apprentissage peut être illustré comme suivant.

### Bon apprentissage vs sur-apprentissage

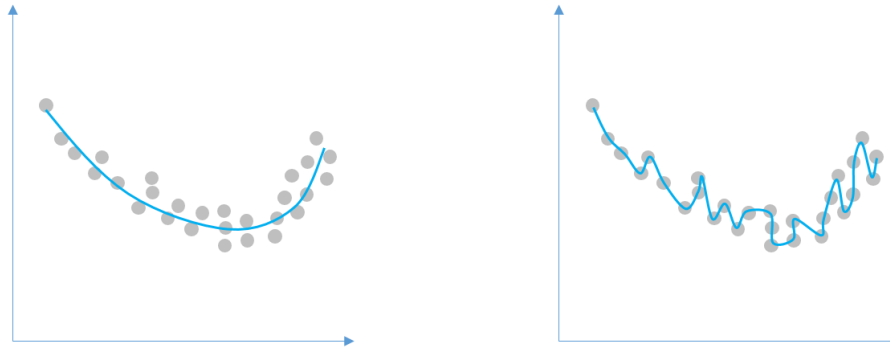


Figure 21 - Représentation du sur-apprentissage

Finalement, utiliser l'apprentissage supervisé dans le cadre de cette étude est pertinent car nous aurons la variable à expliquer  $Y$ , qui pourra correspondre à la PVFP ou au SCR, et qui sera disponible dans notre base de données. Nous traiterons ainsi par la suite seulement des modèles d'apprentissage supervisé et plus particulièrement ceux permettant de traiter un problème de régression.

## 2.4 Présentation des différents modèles

De nombreux modèles de *machine learning* permettent de traiter les problèmes de régression. Nous verrons ci-dessous le principe de différents modèles utilisés dans ce contexte. Afin de comparer la pertinence de ces méthodes de machine learning, le modèle linéaire généralisé sera aussi présenté. En effet, c'est un modèle très répandu en actuariat, qui pourrait notamment permettre de répondre à notre problématique.

### 2.4.1 Modèle linéaire généralisé (GLM)

Ce modèle est une extension du modèle de régression linéaire classique. L'objectif est d'établir une relation entre une variable à expliquer  $Y$  et des variables explicatives  $X = \{X_1, X_2, \dots, X_p\}$ . Les GLM sont ainsi constitués de 3 composantes :

- $Y$  qui est une composante aléatoire
- La combinaison linéaire des variables  $X = \{X_1, X_2, \dots, X_p\}$  qui est une composante déterministe, aussi appelée prédicteur linéaire
- La fonction de lien  $g$  qui définit la nature de la relation entre la composante déterministe et l'espérance de la composante aléatoire

Le modèle peut ainsi être écrit de la façon suivante :

$$g(E[Y]) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = \beta_0 + \sum_{i=1}^p \beta_i X_i$$

#### La composante aléatoire

L'espérance de  $Y$  doit être déterminée à partir de  $n$  réalisations  $\{y_1, \dots, y_n\}$  indépendantes de la loi de  $Y$ . La loi de probabilité de cette composante doit appartenir à la famille exponentielle, c'est-à-dire que qu'elle peut être écrite sous la forme :

$$f(y, \theta, \varphi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi) \right\}$$

Avec  $\theta \in \mathbb{R}$  le paramètre de la moyenne (ou paramètre canonique),  $\varphi \in \mathbb{R}$  le paramètre de dispersion (ou paramètre de nuisance) et les fonctions  $a$ ,  $b$  et  $c$  tel que :

- $a : \mathbb{R} \rightarrow \mathbb{R}^*$
- $b : \mathbb{R} \rightarrow \mathbb{R}$  de classe  $C^2$
- $c : \mathbb{R} \rightarrow \mathbb{R}^2$

Ainsi différentes lois sont éligibles en fonction du problème à traiter, comme les lois Normale, de Poisson, Binomiale, Gamma, inverse Gaussienne, Bernoulli, ... Nous pouvons prendre l'exemple des lois suivantes, souvent associées à différents problèmes :

- Pour un résultat continu et illimité, la distribution Normale
- Pour un résultat continu et non négatif, la distribution Gamma ou inverse Gaussienne
- Pour un résultat discret, la distribution de Poisson. Cette a néanmoins une limite ; Il est supposé que la moyenne soit égale à la variance. Si cette condition n'est pas remplie, il est possible de considérer la distribution quasi-Poisson ou la distribution Binomiale négative
- Pour un résultat binaire ou de proportions, la distribution Binomiale.

### La composante déterministe

Pour chaque réalisation  $\{y_1, \dots, y_n\}$  nous disposons de  $p$  variables explicatives associées  $\{x_1, x_2, \dots, x_p\}$ . Nous obtenons ainsi le prédicteur linéaire de la forme :

$$\beta_0 + \sum_{i=1}^p \beta_i x_i$$

Avec  $\beta$  des coefficients de régression.

Il est possible de rajouter des combinaisons de variables explicatives dans le modèle comme  $X_i \times X_j$  qui correspond à l'interaction entre les variables  $X_i$  et  $X_j$ . Il est aussi possible de prendre en compte l'effet non-linéaire d'une variable  $X_i$  en la mettant sous la forme  $X_i^2$ . Toutes les variables ne seront pas forcément utilisées dans la GLM, seules les variables significatives seront retenues.

### La fonction de lien

La fonction  $g$  est différentiable et monotone, et permet de relier l'espérance de  $Y$  au paramètre canonique  $\theta$ . On a ainsi :

$$g(E[Y]) = \beta_0 + \sum_{i=1}^p \beta_i X_i = \theta$$

A chaque loi de probabilité de la famille exponentielle est associée une fonction de lien canonique. Il faudra étudier les données afin de définir laquelle sera la plus pertinente.

### Sélection des variables

Utiliser un modèle comprenant des variables ayant un certain lien entre elles ou peu significatives peut entraîner un biais, il est donc important de ne garder que les variables pertinentes. Afin de pouvoir les comparer et sélectionner celles qui expliquent au mieux la variable  $Y$ , plusieurs démarches peuvent être utilisées :

- La méthode descendante, qui consiste à implémenter le modèle tout d'abord avec l'ensemble des variables. La variable la moins significative sera ensuite enlevée du modèle jusqu'à un certain seuil. La significativité des variables pourra être définie par son effet sur le  $R^2$  ou selon des critères précisés dans la partie suivante.
- La méthode ascendante, qui fonctionne de façon inverse à la méthode descendante. Le premier modèle comprend donc seulement le terme constant  $\beta_0$  et les variables sont ensuite intégrées les unes après les autres.
- La méthode progressive, qui mélange les deux démarches ci-dessus.

### Comparaison des modèles

Plusieurs critères peuvent être utilisés afin de comparer différents modèles. On peut tout d'abord utiliser la déviance qui s'obtient en comparant le modèle estimé au modèle « saturé », c'est-à-dire le modèle qui possède autant de paramètres que d'observations. Le modèle saturé représente donc exactement les données. Ainsi, la déviance vaut :

$$D = -2(L - L_{sat})$$

Avec  $L$  la log-vraisemblance du modèle estimé et  $L_{sat}$  la log-vraisemblance du modèle saturé. Une déviance faible correspond à une bonne qualité d'ajustement, l'objectif est donc de minimiser ce critère.

Un autre critère très utilisé est le critère d'information d'Akaike noté AIC. Il prend en compte la qualité d'ajustement car il est composé de la fonction de vraisemblance, mais aussi de la complexité du modèle car il prend en compte le nombre de paramètres intégrés. Il s'écrit de la façon suivante.

$$AIC = -2L + 2k$$

Avec  $k$  le nombre de paramètres du modèle.

Le critère d'information bayésienne BIC est une amélioration du critère AIC qui permet de prendre en compte le nombre d'observations  $n$ . Il s'écrit comme suivant.

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

## 2.4.2 Arbres de décision

Les arbres de décision CART (*Classification And Regression Trees*) font partie des méthodes d'apprentissage statistique permettant de résoudre des problèmes de classification et de régression. C'est un modèle non-paramétrique, ce qui implique que son utilisation ne nécessite pas d'hypothèse de loi en amont. L'objectif de l'arbre de décision sera d'expliquer une variable  $Y$  à partir de variables explicatives  $X = \{X_1, X_2, \dots, X_p\}$  en effectuant une segmentation des données selon certaines règles.

Le nom « arbre de décision » provient de la représentation visuelle de l'algorithme. L'ensemble des individus correspond à la racine de l'arbre. Les individus seront ensuite scindés en deux sous-ensembles en fonction d'une règle admettant une réponse binaire. Les sous-ensembles formeront des nœuds où de nouvelles séparations pourront s'effectuer. Les dernières séparations formeront les feuilles de l'arbre, et regrouperont donc les individus ayant des caractéristiques similaires. Un arbre de décision peut être représenté de la façon suivante.



## Structure d'un CART

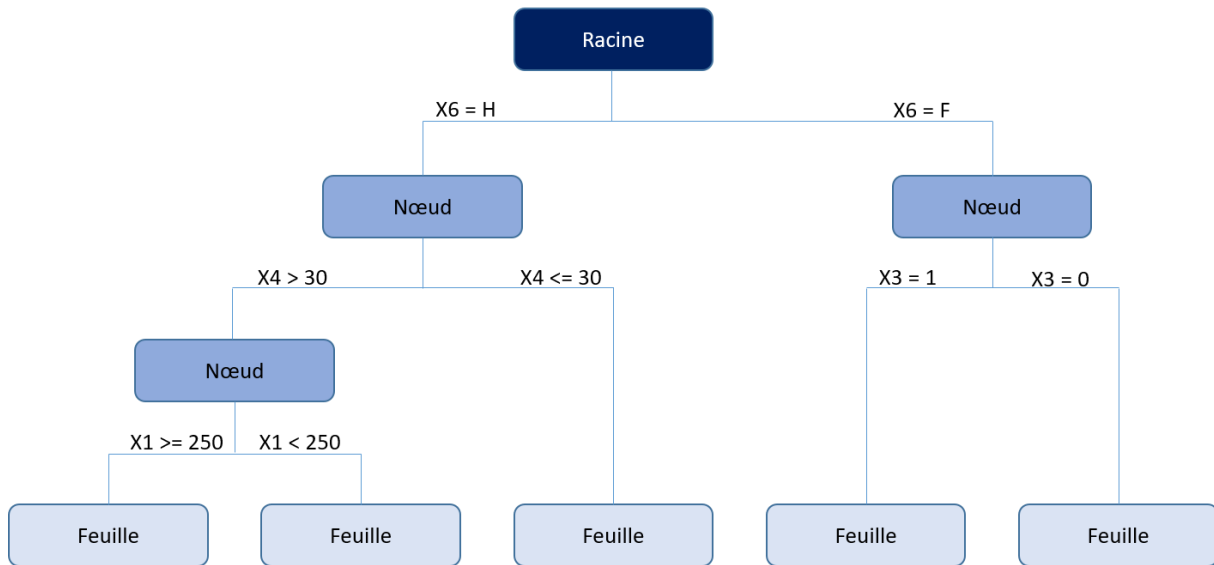


Figure 22 - Représentation de la structure d'un arbre de décision

Dans cette illustration, les individus sont tout d'abord séparés en fonction de la valeur de la variable  $X_6$ . Nous retrouverons dans le nœud à droite les individus ayant comme valeur « F » pour la variable  $X_6$ . Ce sous-ensemble sera encore une fois scindé en fonction des valeurs associées à la variable  $X_3$  ; Les individus pour lesquels  $X_3 = 1$  iront dans la feuille de gauche et ceux dont  $X_3 = 0$  iront dans celle de droite. Finalement, on peut observer que les individus comprenant des caractéristiques similaires se trouvent regroupés dans une même feuille.

La construction de l'arbre de décision se fait en deux étapes. Tout d'abord la création de l'arbre maximal sur une base d'apprentissage est réalisée, puis l'élagage de cet arbre et la construction de l'arbre optimal sont effectués.

### Construction de l'arbre maximal

Soit  $\pi_0$  la quantité que l'on veut prédire définie par  $\pi_0 = E[Y|X = x]$ . Le but est d'expliquer  $\pi_0$  en se plaçant dans un cadre de régression. Nous cherchons à avoir un partitionnement qui maximise l'homogénéité dans les classes construites.

Etant donné que nous voulons prédire une espérance, la fonction de perte que nous pouvons utiliser est l'erreur quadratique moyenne (MSE) tel quel :

$$MSE = E[(\pi(x) - Y)^2]$$

Le critère de division doit ainsi être minimisé. La solution de  $\pi_0$  est obtenue par :

$$\pi_0 = \arg \min_{\pi(x)} E[\phi(Y, \pi(x)) | X = x]$$

Avec  $\phi(Y, \pi(x)) = (\pi(x) - Y)^2$

Il est à noter que dans le cas d'un problème de classification, le critère de division est différent. Il s'agit d'une minimisation de l'impureté d'un nœud que l'on appelle critère de Gini. Les variables réponses que nous allons considérer étant quantitatives, ce critère ne sera pas utilisé.

Lors de la construction de l'arbre, une segmentation est effectuée pour chaque nœuds afin d'obtenir deux nœuds « fils ». Les nouveaux nœuds seront aussi segmentés jusqu'à ce qu'intervienne notre critère d'arrêt.

Afin de s'assurer d'avoir une séparation des nœuds en sous-ensembles homogènes, chaque variable explicative et chaque seuil sont testés. L'algorithme retient finalement la combinaison qui minimise la somme des variances des deux sous-ensembles, ce qui crée ainsi deux sous-ensembles les plus disjoints possibles.

### Elagage de l'arbre

L'arbre maximal construit est soumis à un grand risque de sur-apprentissage étant donné que le nombre de feuilles produites peut être très élevé. Le modèle est ainsi dans l'impossibilité de généraliser ses résultats sur une base de test, et cela limite donc son utilisation. L'objectif de l'élagage est de réduire la complexité de l'arbre afin de palier à ce problème. Cela peut être effectué de deux façons :

- Lors de la construction d'un arbre directement, c'est-à-dire en rajoutant des conditions comme le nombre d'individus minimal par feuille. Il est aussi possible de définir la profondeur de l'arbre, soit le nombre maximal de variables considérées pour construire l'arbre.
- A partir de l'arbre maximal, en supprimant les nœuds les plus bas de l'arbre s'ils n'apportent qu'un gain minimal sur la variance. C'est une approche par coût-complexité, où un paramètre de complexité  $\alpha$  est intégré. L'objectif est de calibrer  $\alpha$  afin d'obtenir un sous-arbre robuste.

### Avantages et limites du modèle

Tout d'abord, les arbres de décision ont une phase de traitement des données plutôt simple. Le type de variables que ce soit quantitatif ou qualitatif n'a pas d'impact significatif lors de la construction de l'arbre. Il n'y a pas besoin de transformer par exemple les variables quantitatives en variables qualitatives, l'algorithme s'adaptera dans tous les cas.

La représentation du modèle sous forme d'arbre rend son interprétation facile. Cela distingue les arbres des autres modèles de *machine learning* dont la compréhension des résultats n'est pas toujours aisée.

Néanmoins par leur construction, les arbres sont très sensibles au risque de sur-apprentissage. L'élagage de l'arbre est donc une étape primordiale afin de limiter ce risque. Les arbres sont aussi très instables et leur performance va fortement dépendre des individus compris dans la base d'apprentissage.

Afin de palier à ces limites, d'autres modèles ont été mis en place en héritant des propriétés intéressantes des CART. Deux axes sont principalement adressés dans ces nouveaux modèles, la perturbation aléatoire des arbres ainsi que la combinaison de plusieurs arbres au lieu de ne se baser que sur un.

### 2.4.3 Forêts aléatoires

Le modèle de forêts aléatoires, ou *Random Forest*, est une approche d'amélioration des CART qui est basée sur un principe basique détaillé dans ce qui suit. La construction d'un seul arbre complexe étant sujette à différentes limites comme le risque de sur-apprentissage ou une grande instabilité, le principe est de former un grand nombre d'arbres plus simples qui permettront ensemble de réaliser une prédiction plus fiable. L'agrégation des différents arbres apporte une certaine robustesse au modèle tout en gardant les aspects attractifs des CART comme leur paramétrage généralement simple par rapport à d'autres modèles.

Représentation d'une prédiction de  $Y$  avec une forêt aléatoire

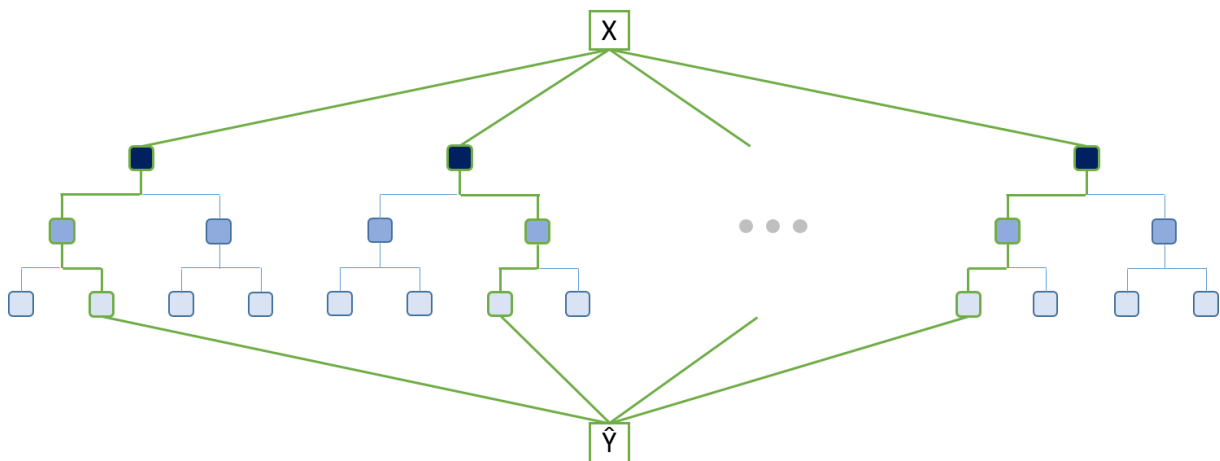


Figure 23 - Représentation du processus d'une forêt aléatoire

Afin d'effectuer une prédiction sur une variable à expliquer  $Y$ , la forêt aléatoire prend en entrée un vecteur de variable explicative  $X$ . Les chemins en vert correspondent au regroupement de l'individu dans la feuille de l'arbre qui lui correspond. Chaque arbre aura donc une estimation de  $Y$  qui lui est propre. Enfin, l'estimation finale de  $Y$  par la forêt va correspondre dans le cas d'un problème de régression à la moyenne des estimations de  $Y$  de chacun des arbres. Pour un problème de classification, le choix final de la valeur de  $Y$  sera issu d'un vote par les arbres, l'estimation la plus fréquente dans la forêt sera la valeur retenue.

Les forêts aléatoires sont réalisées à l'aide de deux algorithmes qu'il est important d'expliciter, c'est le *tree bagging* ainsi que le *feature sampling*.

#### Tree bagging

C'est une technique de ré-échantillonnage que l'on appelle aussi *bootstrap aggregating*, et qui consiste à sélectionner aléatoirement pour la construction de chaque arbre un sous-ensemble d'individus avec remise. Ce sous-ensemble constituera la base d'apprentissage de l'arbre, et l'agrégation des différents arbres construits de cette façon constitue le *tree bagging*.

Ainsi, cette méthode permet de s'assurer que la forêt soit constituée d'arbres différents. Plus le nombre d'arbres construits dans la forêt est grand, plus robuste sera le modèle, mais le temps de calcul nécessaire augmentera tout autant.

### Feature sampling

Cela consiste en l'intégration d'un aléa sur les variables considérées dans la construction des arbres. En effet, une limite du *bagging* est que lorsque certaines variables ont un pouvoir explicatif très grand, les arbres ainsi construits ont un risque plus important d'être corrélés entre eux. Ainsi, lors de la génération d'un nouveau nœud, seul un sous-ensemble aléatoire de variables explicatives est considéré afin de limiter cet effet.

La taille du sous-espace est un paramètre qui doit être calibré lors de la création d'une forêt aléatoire. Des valeurs empiriques sont néanmoins proposées par défaut selon le type de problème. En effet pour  $p$  variables explicatives, la taille du sous-espace suggéré est de  $\sqrt{p}$  pour un problème de classification, et  $\frac{p}{3}$  pour un problème de régression. L'utilisation d'un algorithme de validation croisée permettra néanmoins de s'assurer de la pertinence de ces paramètres.

En effet, un algorithme de validation croisée à  $k$  blocs consistera à diviser la base de données en  $k$  échantillons. Il s'agira par la suite de sélectionner un bloc comme ensemble de validation, et les  $k - 1$  blocs restants constitueront la base d'apprentissage. Après entraînement, la performance du modèle sera calculée sur l'échantillon de validation. L'opération sera répétée en tout  $k$  fois, c'est à dire avec chaque bloc défini. La performance globale tiendra compte de la moyenne et de l'écart-type des  $k$  scores de performance.

### Avantages et limites du modèle

Les composantes aléatoires du modèle permettent de limiter les risques de sur-apprentissage, ce qui constitue la majeure amélioration par rapport aux CART, algorithmes sur lesquels les forêts aléatoires sont basées. De plus, c'est l'un des meilleurs modèles de *machine learning* afin d'obtenir des prédictions précises.

Néanmoins les forêts aléatoires montrent deux limites en pratique. Tout d'abord d'un point de vue computationnel, même si les arbres sont générés en utilisant une parallélisation des tâches, le temps de calcul est assez long lorsque l'on crée beaucoup d'arbres ou que la base de données est de grande dimension. De plus, la prédiction de variables extrêmes aura des résultats instables dans le cadre d'un problème de régression.

Enfin, il est adéquat de mentionner l'importance de l'explicabilité du modèle dans un cadre assurantiel. La mise en relation des valeurs prises par certaines variables et leur conséquence sur la prédiction est un enjeu dans différents domaines de l'actuariat. Les forêts aléatoires sont ainsi limitées par leur explicabilité, et souvent qualifiées de modèle « boîte noire ».

### 2.4.4 Extreme gradient boosting (XGboost)

Ce modèle est une autre approche d'amélioration des CART. Les forêts aléatoires répondent aux contraintes des arbres de décision en utilisant le *tree bagging*. Ainsi, les arbres réalisés sont indépendants les uns des autres et la prédiction va correspondre à la moyenne des prédictions de tous les arbres. XGboost est une variante du modèle GBM (*Gradient Boosting Model*) qui est basé sur une méthode différente : le *boosting*.

## Boosting

C'est une méthode d'agrégation qui repose sur une stratégie d'optimisation des poids par récurrence. Elle peut être utilisée pour différents types de prédicteurs, néanmoins dans le cadre de ce mémoire, son fonctionnement avec des arbres de décision sera considéré.

Le *boosting* construit les arbres en série par contradiction aux forêts. Chacun des arbres générés aura ainsi l'information de l'erreur de l'arbre précédent. Afin de rendre le prédicteur plus performant, le poids associé aux données comprenant l'erreur la plus importante est augmenté. Chaque itération cherchera donc à améliorer la qualité prédictive du modèle précédent en se focalisant principalement sur ses lacunes. Les illustrations ci-dessous montrent les approches des algorithmes CART, forêt aléatoire (*bagging*) et XGboost (*boosting*) par rapport aux observations considérées.

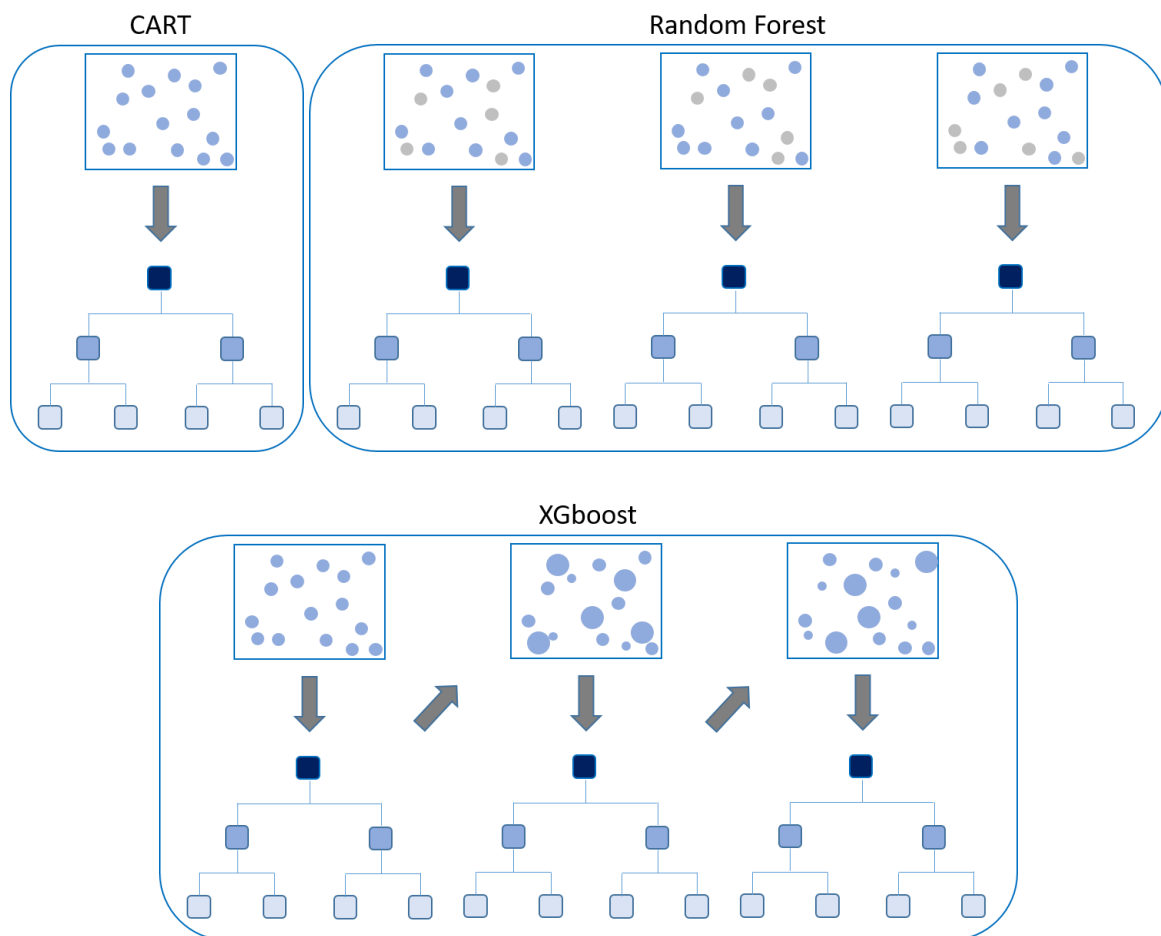


Figure 24 - Approche des algorithmes CART, random forest et XGboost par rapport aux observations

Les CART sont construits en prenant en compte toutes les observations. Pour les forêts aléatoires, chaque arbre est formé en parallèle avec seulement un sous-échantillon des données en entrée. Pour XGboost, le premier arbre est généré avec l'ensemble des données comme pour un CART, puis les suivants sont construits en séquence à partir des erreurs commises en adaptant avant chaque nouvel arbre le poids des observations. Le poids augmentera en fonction des lacunes du modèle.

Une fonction de coût est utilisée afin d'optimiser les estimateurs. Un avantage du *boosting* est le libre choix de cette dernière, néanmoins les modèles GBM et XGboost utilisent la méthode de descente du gradient. La combinaison du *boosting* et de descente du gradient est particulièrement performante, elle est appelée *gradient boosting* et constitue la base de ces modèles.

### Avantages et limites du modèle

XGboost apporte néanmoins des avantages par rapport au GBM. Tout d'abord, son implémentation utilise une parallélisation, afin de réduire fortement les temps de calculs. De plus le prédicteur utilisé n'est pas forcément un arbre de régression contrairement au GBM. De façon plus générale, GBM et XGboost apportent tous deux une solution alternative aux forêts aléatoires dont les performances sont généralement très intéressantes en dépit d'un paramétrage plus difficile.

Bien que les CART soient marqués par leur facilité d'interprétation, ses améliorations basées sur le boosting entraînent une difficile explicabilité. La problématique de modèle « boîte noire » limite donc l'utilisation de XGboost dans différents domaines.

Dans la suite du mémoire, les modèles présentés ci-dessus seront utilisés afin d'évaluer leur pertinence dans le cadre de la détermination d'une allocation optimale. Nous choisissons d'utiliser une base de données comprenant les valeurs de la PVFP et du SCR de marché. Ces valeurs ont été calculées notamment en fonction de différentes allocations d'actifs. Cette base de données sera ainsi générée en utilisant le modèle ALM interne à Optimind sur un portefeuille d'épargne fictif.

## 2.5 Généralités d'un contrat d'épargne, modèle ALM et GSE utilisés

Le portefeuille sur lequel sera basée cette étude sera un portefeuille d'épargne. Nous rappellerons dans la section suivante quelques généralités qui concernent ce produit. Une description du fonctionnement du modèle ALM et du GSE utilisés sera ensuite réalisée.

### 2.5.1 Généralités du contrat d'épargne

Le contrat d'épargne est un contrat d'assurance vie qui permet à un assuré de verser des primes à un assureur afin de percevoir un capital ou une rente à partir d'une certaine date contractuelle. Il se caractérise par plusieurs éléments.

#### Le type de versement

Les primes, aussi appelées versements ou cotisations, peuvent être de 3 catégories différentes :

- Périodiques : ce sont des versements à durée, période et montant prédéfini au début du contrat. En général, avec cette modalité, les premiers versements sont prélevés par l'assurance afin de payer la totalité des frais du contrat.
- Libres : le choix du montant et moment de cotisation est laissé à la discrétion de l'assuré.
- Uniques : un seul versement est autorisé au moment de la souscription du contrat.

#### Le type de support

Il existe 2 principaux types de supports pour le contrat d'épargne :

- Le contrat mono-support : aussi appelé contrat en euro, le principe est que le risque financier est porté partiellement par l'assureur. Dans ce cas, l'assuré définit contractuellement avec son assureur le taux minimum garanti (TMG) appliqué à son épargne. De plus, les produits financiers dégagés par les fonds doivent être reversés aux assurés. En effet, au moins 85% du résultat financier et 90% du résultat technique sont reversés immédiatement ou bien dans les 8 ans. L'assureur est considéré comme le porteur de risque car il garantit un capital minimum à une date donnée quel que soit le résultat de ses placements financiers.
- Le contrat multi-support : c'est un contrat qui contient au moins une partie investie en unité de compte (UC). Ici le risque financier est porté par l'assuré. L'assuré possède un nombre de parts d'actif en fonction du montant de la prime et de la valeur liquidative du support. Ce nombre de parts ne change pas mais la valeur du support change en fonction du temps. L'intérêt de ce support est que le risque et le rendement sont plus élevés qu'un contrat en euro, et peut donc correspondre à des profils d'investisseurs différents.

D'autres types de contrats existent tels que l'euro-croissance ou bien NSK, ceux-ci se rapprochent du mode de fonctionnement des contrats multi-supports mais avec une fiscalité avantageuse sous certaines conditions.

## Le type de chargement

Les chargements sont les frais prélevés par l'assureur (par exemple lors de prestations) et qui ont pour but de couvrir ses dépenses lors de la gestion des contrats. Il existe de nombreux types de chargements :

- Frais de dossier : c'est un montant fixe à payer une seule fois lors de la signature du contrat. Ce chargement n'est pas présent sur tous les contrats.
- Chargements sur prime : aussi appelés frais sur versement pour l'assuré, frais de souscription ou chargement d'acquisition, ils permettent de couvrir les frais d'acquisition. Ces frais correspondent à un pourcentage des primes versées et sont prélevés en général à chaque cotisation.
- Chargements de gestion : aussi appelés chargements sur encours, ce sont des frais prélevés périodiquement sur la vie du contrat. Ils sont appliqués au montant des primes payées sur la période ou de la provision mathématique.
- Chargements de fractionnement : cela permet de couvrir les frais bancaires ou de gestion liés aux prélèvements bancaires. C'est un montant fixe ou un taux prélevé à chaque versement, en général un taux dégressif en fonction du montant de la prime.

## Garanties et prestations

L'assureur propose à ses assurés plusieurs prestations ainsi que garanties sur son contrat d'épargne :

- Garantie épargne : en cas de vie de l'assuré, l'assureur reverse au terme du contrat un capital ou une rente au bénéficiaire. Cette garantie est obligatoire dans chaque contrat.
- Garantie décès : en cas de décès de l'assuré, l'assureur reverse un capital ou une rente au bénéficiaire. Le montant peut varier en fonction des options choisies pour le contrat. La garantie décès ouvre à une fiscalité avantageuse pour les successions dans un certain cadre.
- Rachat : pendant la durée du contrat l'assuré peut reprendre une partie de son épargne. Plusieurs options s'offrent à lui :
  - Rachat total : dans ce cas toute l'épargne est reprise avant le terme du contrat. Dans le cas de fonds en euro, l'assuré reçoit son épargne déduite d'éventuelles pénalités, et pour les fonds en UC, il reçoit la valeur en euro des actifs qu'il possède.
  - Rachat partiel : le fonctionnement est identique au rachat total, la seule différence est qu'une partie de l'épargne reste investie.
  - Avance : cela permet à l'assuré de reprendre une partie de son épargne temporairement. Le fonctionnement est proche de celui d'un prêt avec taux d'intérêt auprès de l'assurance. Dans le cas où il ne rembourserait pas, le capital ainsi que ses intérêts sont prélevés de son épargne.
- Arbitrage : pour les contrats en UC, c'est la possibilité pour l'assuré de transférer ses parts d'actifs sur un support différent.
- Taux Minimum Garanti (TMG) : ce taux est défini contractuellement avec l'assuré et concerne les fonds en euros. Il garantit à l'assuré un rendement minimum de son épargne (ou nul) et peut être redéfini chaque année. Légalement, il ne peut pas excéder « 80 % du produit de la moyenne des taux de rendement des actifs de l'entreprise calculée pour les deux derniers exercices, par les provisions mathématiques des contrats » (Code des assurances Article A132-3 Alinéa 1).



En dehors de ces éléments contractuels, plusieurs mécanismes ont une place importante dans les contrats d'épargne. C'est notamment le cas de l'effet cliquet qui désigne le fait que les intérêts gagnés pour l'assuré avec son épargne, sont acquis définitivement. Cet effet ne concerne que les fonds investis en euros, car les assureurs utilisent ce capital pour effectuer des placements principalement de type obligataire. L'assureur a ainsi un rendement plus faible mais aussi un risque associé plus faible.

Un autre mécanisme, socle de l'épargne en assurance vie, est la Participation aux Bénéfices (PB) qui concerne elle aussi les fonds en euros seulement. C'est la somme des bénéfices financiers et techniques effectués au cours de l'année. Un cadre légal impose aux assureurs de redistribuer au minimum 85% du résultat financier et 90% du résultat technique, aux assurés immédiatement, ou dans un délai de 8 années maximum. Dans le cas où l'assureur choisit de conserver la PB pendant quelques années, il enregistre celle-ci en Provision pour Participation aux Bénéfices (PPB). La provision permet aux assureurs de lisser leur résultat sur plusieurs années afin de mieux satisfaire leurs assurés et actionnaires, et ne peuvent effectuer de reprise que dans le cas d'une perte annuelle.

## 2.5.2 Modèle ALM

L'outil de projection qui a été utilisé est le modèle ALM épargne d'Optimind, développé sous Excel et VBA. Il permet d'effectuer une simulation de l'actif et du passif sur un horizon de 40 ans pour valoriser le coût des garanties et options, et d'évaluer le *Best Estimate* de façon déterministe ou stochastique. Ainsi, il projettera notamment les différentes classes d'actifs, et effectuera un vieillissement du passif en tenant compte des différentes interactions actif-passif.

Pour le calcul du SCR, l'approche utilisée est la formule standard. L'outil permettra d'obtenir à l'issue des projections différentes métriques, telles que la PVFP ou le SCR dont nous avons besoin dans le cadre de ce mémoire. Il est à noter que l'évaluation sous Solvabilité II du modèle suit une hypothèse de *run-off*. En ce sens, il n'y aura pas de nouveaux contrats souscrits pendant la projection.

Nous verrons dans la suite de ce mémoire le fonctionnement du modèle dans un premier temps, puis les mécanismes de modélisation du passif et de l'actif ainsi que le mécanisme de la participation aux bénéfices.

### Fonctionnement du modèle

Le modèle est alimenté à partir des tables de scénarios économiques produites du GSE qui sera présenté dans la suite du mémoire. L'actif et le passif du portefeuille seront projetés sur un horizon défini à 40 ans pour chaque scénario économique. Les tables du GSE comprennent 1000 simulations. Le schéma suivant permet de représenter l'architecture de l'outil :

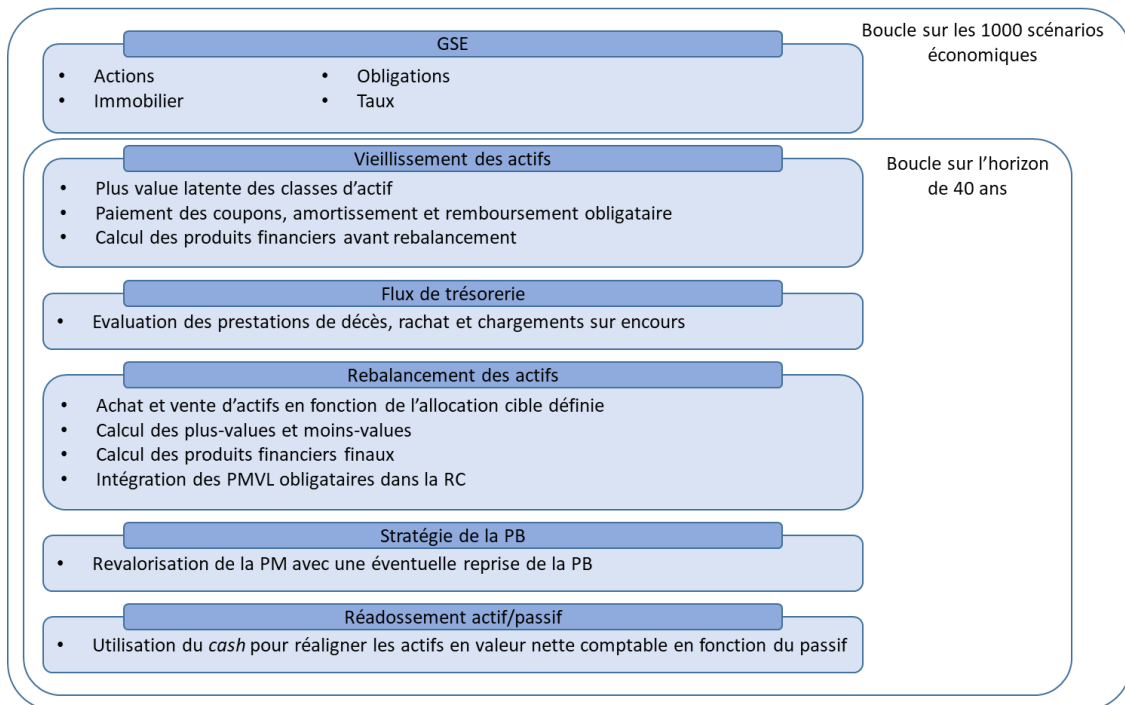


Figure 25 - Architecture du modèle ALM inspiré d'une note technique Optimind

### Modélisation du passif

Le passif est structuré en regroupements d'assurés avec des caractéristiques similaires, aussi appelés *model points*. Ces derniers ont été construits à partir de données d'un portefeuille fictif, qui permet de représenter un assureur typique selon l'observation d'experts. Le modèle comprend les provisions suivantes :

- La PM (Provision Mathématique) : définie par la « différence entre les valeurs actuelles des engagements respectivement pris par l'assureur et par les assurés » selon l'article R331-3 du Code des Assurances. C'est le montant nécessaire à l'assureur pour répondre à ses engagements vis-à-vis des assurés. Dans le cas d'un contrat d'épargne, cela correspond à la somme des primes nettes versées par les assurés, revalorisée chaque année des intérêts calculés en fonction de la TMG du contrat et de la PB (Participation aux Bénéfices) reçue.
- La PPB (Provision pour Participation aux Bénéfices) : aussi appelée PPE (Provision pour Participation aux Excédents), correspond à la somme des bénéfices de l'assureur due aux assurés. Ici, le mécanisme de distribution de la PB employé est la distribution de la PB contractuelle, complétée par une éventuelle utilisation de la PPB afin de servir le taux cible.
- La RC (Réserve de Capitalisation) : elle permet de lisser le résultat de l'assureur sur les titres obligataires. En effet, lors de la cession d'obligation, deux scénarios s'opposent :
  - o Si le prix de vente est inférieur à la valeur nette comptable, la plus-value vient augmenter la RC
  - o Si le prix de vente est supérieur à la valeur nette comptable, la plus-value vient diminuer la RC

Les prestations sont estimées dans l'outil de la façon suivante : dans un premier temps les décès, puis les rachats.

### **Décès**

La mortalité est modélisée à partir des tables fournies par l'INSEE « TF 00-02 » pour les femmes et « TH 00-02 » pour les hommes. Cette prestation correspond à la PM d'un assuré et est immédiatement payée par l'assureur.

## Rachats

Le modèle épargne ne distingue pas les rachats partiels des rachats totaux. Un taux de rachat est appliqué sur les *models points* pour obtenir la somme des rachats par année de simulation. Etant donné que ces derniers peuvent être dus à un choix de l'assuré ou à une conjoncture économique, l'outil permet de modéliser ces deux composantes séparément. Le taux de rachat correspond à la somme des rachats structurels et conjoncturels.

Les rachats structurels correspondent aux rachats survenant dans un environnement économique considéré comme normal. Il va dépendre de différents facteurs, néanmoins en assurance vie il est fortement lié à la fiscalité avantageuse. En effet, étant donné une forte baisse du taux d'imposition à partir de 8 ans, le taux de rachat observé est en croissance jusqu'à ce palier atteint. Il est ensuite constant. Cette hypothèse de taux est un paramètre du modèle.

Les rachats conjoncturels quant à eux dépendent de l'environnement économique, et plus particulièrement du taux servi par l'assureur. En effet, lorsque le taux servi par son assurance ne lui convient pas, l'assuré risque de procéder à un rachat de son contrat pour aller chez un autre assureur plus compétitif. Le taux dépend donc de l'écart entre le taux servi et le taux attendu. Le modèle ALM utilise la loi préconisée par l'ACPR :

$$RC = \begin{cases} RC_{max} & si TS - TA < \alpha \\ RC_{max} \frac{(TS - TA - \beta)}{\alpha - \beta} & si \alpha \leq TS - TA < \beta \\ 0 & si \beta \leq TS - TA < \gamma \\ RC_{min} \frac{(TS - TA - \gamma)}{\delta - \gamma} & si \gamma \leq TS - TA < \delta \\ RC_{min} & si TS - TA \geq \delta \end{cases}$$

Avec :

- $RC_{max}$  le taux de rachats conjoncturels maximum et  $RC_{min}$  le taux minimum
- TS le taux servi
- TA le taux attendu
- $\alpha$  le seuil à la hausse à partir duquel le taux de rachats conjoncturels est considéré comme constant
- $\beta$  et  $\gamma$  les seuils d'indifférence entre lesquels le taux de rachat conjoncturel est nul
- $\delta$  le seuil à la baisse à partir duquel le taux de rachats conjoncturels est considéré comme constant

## Modélisation de l'actif

Différentes classes d'actifs sont présentes dans le modèle :

- Les actions
- L'immobilier
- Les obligations
- Le *cash*

## Action et immobilier

On distingue les actions de type 1 et les actions de type 2. Dans le cadre de ce mémoire, seules les actions de type 1 ont été modélisées. Pour rappel, elles correspondent aux actions listées sur un marché réglementé ou échangées sur une plateforme multilatérale de négociation dans un pays membre de l'UE ou de l'OCDE.

Le modèle épargne permet de paramétrer différentes hypothèses sur ces actifs :

- L'allocation cible, soit la proportion d'actions ou d'immobilier souhaitée dans le portefeuille.
- Le taux de plus-value initiale, qui permet d'obtenir la valeur de marché à partir de la valeur nette comptable.
- Le taux de plus-value automatique, aussi appelé *turnover*, qui correspond à un pourcentage des actifs vendus lors de la projection, lorsque la valeur de marché est supérieure à la valeur nette comptable. Le taux de plus-value automatique est ainsi appliqué sur l'excédent.

L'évolution du cours des actions et de l'immobilier sont contenus dans des tables générées préalablement par le GSE. 1000 scénarios ont été générés et chacun est utilisé afin d'effectuer une simulation.

Un vieillissement de ces actifs est effectué à la fin de chaque année de la façon suivante :

- Tout d'abord la valeur de marché (VM) est calculée à partir de la valeur nette comptable (VNC) et du taux de plus-value initiale :
$$VM = VNC \times (1 + \text{Taux}_{\text{plus-value initiale}})$$
- A partir de la fin de la première année de projection, la VNC de la seconde année prend la valeur de la VNC issue de la fin de l'année précédente.
- Lorsque qu'une plus-value est observée, un montant correspondant au taux de plus-value automatique vient arbitrer le portefeuille tel que présenté précédemment.
- Pour chaque scénario, la VM suit l'évolution de l'indice issu du GSE.

A l'issue de ce vieillissement, un rebalancement des actifs est réalisé. L'outil calcul l'assiette de rebalancement qui correspond à l'ensemble des actifs, mais aussi une autre spécifique à chaque actif. La valeur cible des actifs est ainsi calculée, ce qui permet de déduire les arbitrages à effectuer afin d'équilibrer le portefeuille en fonction de notre stratégie d'allocation initiale. Le montant de l'allocation cible présenté ci-dessus est alors d'une importance capitale car il sera pris en compte lors du rebalancement des actifs à chaque année de projection.

## Obligations

Cet actif est d'une grande importance dans la composition d'un portefeuille car son allocation représente en général plus des deux tiers des actifs d'un assureur. Deux types d'obligations sont modélisés :

- Les obligations d'état, qui comme leur nom l'indique sont émises par les états et sont prisées pour leur risque de défaut associé très faible. Elles sont caractérisées par le versement d'un coupon et le remboursement du principal à leur terme.
- Les obligations d'entreprises, qui sont émises par les entreprises et comprennent un risque de défaut plus important associé à un taux plus élevé.

Le modèle épargne permet de paramétrer différentes hypothèses sur cet actif :

- L'allocation cible, soit la proportion d'obligations souhaitées dans le portefeuille.
- L'allocation cible pour chacun des deux types d'obligations.

Le portefeuille d'obligations détenues en situation initiale doit être renseigné dans le modèle. Il est nécessaire de préciser le type de l'obligation, sa maturité, le taux de coupon, la valeur nette comptable et la valeur de marché. Une hypothèse supplémentaire permet de majorer le taux des obligations d'entreprises : c'est le taux de *spread*. La valeur de marché de l'obligation est calculée telle que :

$$VM_m(t) = \sum_{k=1}^m [C_m N_m P(t, k)] + N_m P(t, m)$$

Avec :

- $VM_m(t)$  la valeur de marché de l'obligation de maturité résiduelle  $m$  et à date  $t$ .
- $C_m$  le taux de coupon de maturité résiduelle  $m$ .
- $N_m$  le nominal de maturité résiduelle  $m$ .
- $P(t, k)$  le prix d'une obligation zéro coupon à date  $t$  et de maturité  $k$ , qui vaut :

$$P(t, k) = \exp(-R(t, k)(k - t))$$

Avec :

- $R(t, k)$  le taux d'intérêt pour un emprunt entre les dates  $t$  et  $k$ .

Le vieillissement de la poche obligataire est réalisé de la façon suivante :

- A chaque année de projection, les coupons sont comptabilisés et un recalcul de la VM est effectué.
- Un amortissement linéaire est réalisé dans le but de prendre en compte le remboursement des obligations pour lesquelles la maturité a été atteinte. L'amortissement correspond à la différence entre le prix d'achat de l'obligation et son nominal.

Comme nous l'avons vu pour les actions et l'immobilier, un rebalancement des actifs est ensuite modélisé. La démarche est similaire à celle de ces actifs, avec pour principale différence le fait que les plus et moins-values sont prises en compte pour la dotation de la réserve de capitalisation. Il est à noter que lors de la projection, les obligations acquises sont caractérisées par le fait que le montant du nominal vaut la valeur d'achat de ces dernières.

## Cash

La dernière classe d'actif est le *cash*. Tout comme pour les actifs précédents, il est nécessaire d'indiquer l'allocation cible de la part de *cash* dans le portefeuille. Par ailleurs, lors de la projection cet actif permet de générer des intérêts. A l'issue du rebalancement des autres actifs, le modèle calcule la différence entre le montant de l'assiette qui correspond à l'ensemble des actifs et la somme de toutes les valeurs de marchés. La part de *cash* dans le portefeuille pour l'année de projection suivante correspondra à cet écart.

### Produits financiers

A l'issue de chaque année de projection, les produits financiers sont calculés avec la formule suivante :

$$\text{Produits financiers} = \text{Coupons} + \text{Amortissements} + \text{Intérêts} + \text{PMV rebalancement} + \text{PMV turnover}$$

Avec :

- *Coupons* qui correspond à la somme des coupons perçus dans l'année

- *Amortissements*, l'amortissement comptable des obligations
- *Intérêts* les revenus générés par la part de *cash*
- *PMV rebalancement* le montant des plus ou moins-values réalisées issues du rebalancement des actions et de l'immobilier
- *PMV turnover* qui correspond à la somme des plus ou moins-values automatiques réalisées lors du rebalancement des actions et de l'immobilier

### Modélisation des interactions entre l'actif et le passif

L'évolution de l'actif et du passif implique une interaction entre ces derniers. La politique relative au versement de la participation aux bénéficiés, et donc la revalorisation des contrats, a un impact sur le passif. En effet, elle entraîne l'augmentation des provisions techniques ainsi qu'une modification du comportement des rachats.

### **Participation aux bénéficiés**

Deux types de revalorisations sont prises en compte dans le mécanisme de la participation aux bénéficiés. Tout d'abord la revalorisation cible car l'assureur doit proposer un taux de revalorisation cohérent avec le marché pour éviter des rachats conjoncturels. Ainsi, le modèle tient compte d'un taux de revalorisation cible à atteindre. Ce dernier correspond au taux attendu. Il est à noter que dans le cas d'un *model point* avec un TMG associé supérieur au taux attendu, le taux cible correspondra au TMG.

Ensuite, la revalorisation garantie qui est définie par la revalorisation correspondant aux engagements contractuels de l'assureur envers l'assuré. Elle est dépendante des hypothèses suivantes associées aux *model points* :

$$\text{Revalorisation garantie} = \max (PM_{\text{avant prestation}} \times TMG ; \text{Taux}_{PB \text{ contractuel}} \times \text{Produits financiers})$$

Avec :

- $PM_{\text{avant prestation}}$  le montant de la PM avant prestation du *model point*
- $TMG$  le taux minimum garanti du *model point*
- $\text{Taux}_{PB \text{ contractuel}}$  le taux de participation aux bénéficiés défini pour le *model point*
- $\text{Produits financiers}$  les produits financiers générés sur l'année par le *model point*

Enfin, plusieurs leviers permettent à l'assureur de servir un taux qui atteint le taux de revalorisation cible :

- L'utilisation de la provision pour participation aux bénéficiés
- Les produits financiers de l'année
- Un arbitrage sur les actions et l'immobilier
- La réduction de la marge financière lorsque les autres leviers ne peuvent pas être utilisés

### Calcul de la PVFP et du SCR de marché

Dans notre outil, la PVFP correspond à la valeur actualisée en début de projection du résultat d'exploitation brut selon la formule suivante :

$$\begin{aligned} \text{Résultat d'exploitation brut} = & PM_{\text{ouverture}} - \text{Prestations}_{\text{sans revalorisation décès}} \\ & - \text{Prestations}_{\text{sans revalorisation rachat}} + 15\%(RC) \\ & + PMVL \text{ sur obligation en fin de projection} + \text{Produits financiers} \\ & - \text{Revalorisation de la PM} + \text{Variation PPE} + \text{PV supplémentaires} \\ & - \text{Revalorisation des prestations} \end{aligned}$$

Le calcul du SCR de marché est détaillé précédemment dans le mémoire. Le modèle suit une approche par revalorisation pour le calcul des composants du SCR de marché. Si nous prenons l'exemple du risque de marché, il correspond à la différence entre la NAV avant application du choc immobilier, et la NAV après choc.

### 2.5.3 Générateur de scénarios économiques (GSE)

Comme vu précédemment, le modèle ALM est alimenté par des scénarios économiques issus d'un générateur de scénarios économiques (GSE). C'est un outil important dans notre étude car il va définir la trajectoire d'évolution de nos valeurs financières telles que les taux d'intérêts, ou les rendements. Ces valeurs représentent les facteurs de risques auxquels sont exposés les assureurs, il est ainsi capital de pouvoir simuler ces risques afin d'évaluer de façon cohérente l'évolution de son portefeuille. Les modèles utilisés dans ces générateurs sont liés à l'évolution du contexte économique courant, et vont nécessiter des informations reflétant la situation économique des marchés, à cet instant donné, afin d'être calibrés.

Il est possible d'utiliser 2 types de GSE, le premier est le GSE en univers monde réel. Ce dernier est basé sur un historique de grandeurs financières, l'objectif sera de répliquer l'évolution passée sur notre horizon de projection. La calibration des modèles dans cet univers s'effectue sur des données historiques, dont le choix sera déterminant. La calibration sera aussi sensible à la profondeur de l'historique retenu ainsi qu'au pas du temps. Si l'exemple est pris de l'utilisation d'un indice tel que le S&P500 qui est basé sur 500 grandes entreprises cotées sur la bourse aux Etats-unis, prendre un historique qui remonte jusqu'au début de la crise des *subprimes* donnera un résultat drastiquement différent à un historique remontant jusqu'à la fin de cette crise uniquement. L'aversion au risque des agents économiques est donc prise en compte et les produits financiers sont ainsi plus rémunérateurs par l'application d'une prime de risque sur le rendement.

Le GSE en univers risque neutre propose une approche différente. Ici, le rendement de tous les actifs est en moyenne égal au rendement de l'actif sans risque. L'aversion au risque des agents économiques n'est pas prise en compte dans ces modèles. De ce fait, il n'y a plus d'application d'une prime de risque et le rendement espéré est beaucoup plus faible que dans un univers monde réel. Le régulateur préconise l'utilisation de cet univers théorique dans le contexte de Solvabilité II avec un objectif de valorisation *market-consistent*, c'est-à-dire où les prix sont cohérents avec ceux observés sur le marché. Ainsi, la calibration des modèles implique la minimisation des écarts entre les prix calculés et les prix observés sur le marché.

Plusieurs choix se présentent donc :

- Le modèle utilisé, tels que Hull & White ou Vasicek pour les taux, et Black-Scholes, Heston ou Merton pour l'action et l'immobilier.
- Les données en entrée, utilisées afin de calibrer les modèles, devront notamment être issues d'actifs liquides.
- L'univers de projection, monde réel ou risque neutre.

#### GSE monde réel

La sélection des modèles utilisés dans un GSE monde réel doit respecter certaines contraintes. Nous pouvons notamment mentionner le contexte actuel de taux négatifs, qui implique l'utilisation d'un modèle de taux cohérent avec la situation économique. Les grandeurs que nous cherchons à projeter avec ce générateur sont les actions, l'immobilier, les taux nominaux et l'inflation.

## Action et immobilier

L'action et l'immobilier sont des grandeurs fortement corrélées. Aussi, il est d'usage d'utiliser le même modèle afin de les projeter. Le modèle de projection choisi pour ces actifs est le modèle Black-Scholes à volatilité constante. Pour rappel, ce modèle implique les hypothèses de marché suivantes :

- Le marché est liquide
- Tous les sous-jacents sont divisibles à l'infini
- Les ventes à découvert sont autorisées
- Les emprunts et prêts illimités sont autorisés
- Il n'y a pas de coût de transaction
- Il n'y a pas de dividendes versés

Sous la probabilité historique, ce modèle suit la dynamique suivante :

$$dS_t = S_t(\mu dt + \sigma dW_t)$$

Avec :

- $S_t$  le cours de l'action à l'instant  $t$
- $S_0 = 100$
- $\mu$  le rendement instantané
- $\sigma$  la volatilité du rendement de l'actif risqué
- $W_t$  un mouvement Brownien standard

En appliquant le lemme d'Itô à  $\ln(S_t)$ , on obtient :

$$\ln(S_t) = \ln(S_0) + \left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t$$

Les rendements de  $\ln(S_t)$  suivent ainsi un mouvement Brownien de *drift*  $\left(\mu - \frac{\sigma^2}{2}\right)$  et de coefficient de diffusion  $\sigma$ .

Afin de calibrer ce modèle, les valeurs historiques retenues sont l'Euro Stoxx 50 pour l'action, avec un historique d'avril 2009 à décembre 2020 afin de ne pas prendre en compte la crise des *subprimes*. En effet, la prise en compte d'un historique d'une plus grande profondeur prenant en compte la crise de 2008, voire celle de 2000, entraînerait un rendement qui semblait trop faible par rapport à ce que pouvait espérer un assureur. Pour l'immobilier, l'indice de l'IEIF (Institut de l'Épargne Immobilière et Foncière) immobilier France net a été retenu sur la même période.

Après calibration, 1000 trajectoires ont été générées sur un horizon de 40 ans. Les résultats obtenus en moyenne sont les suivants :



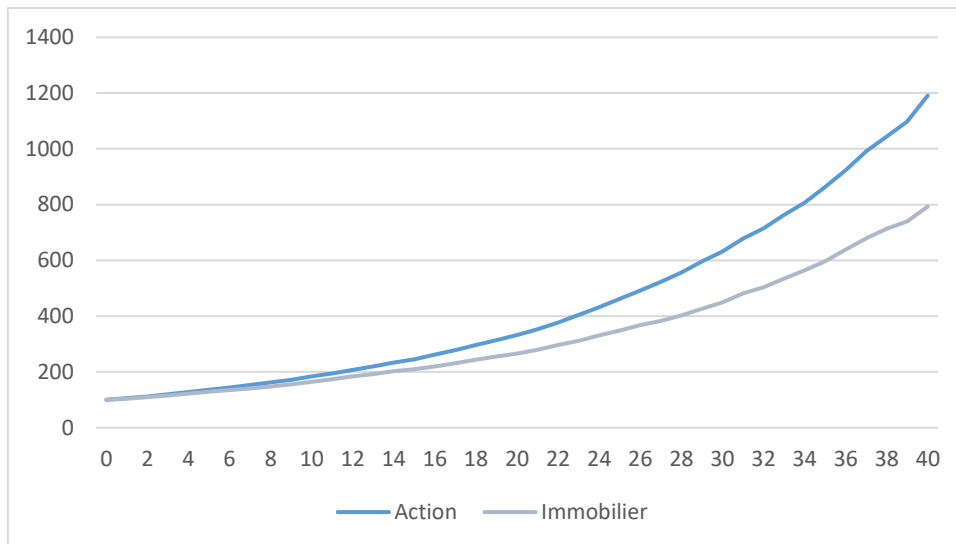


Figure 26 - Projection de l'action et de l'immobilier en moyenne

### Inflation et taux nominaux

Afin de projeter les taux nominaux, le modèle de Hull & White est utilisé car il permet de modéliser des taux négatifs. Par ailleurs, il implique une absence d'opportunité d'arbitrage, ce qui rendrait possible son utilisation dans le cadre d'un GSE risque neutre. L'inflation est couramment projetée avec le même modèle, nous l'utiliserons aussi dans ce cas afin d'être cohérent avec les pratiques du marché.

D'après le modèle de Hull & White à un facteur, le taux court suit la dynamique suivante :

$$dr_t = (b(t) - ar_t)dt + \sigma dW_t$$

Avec :

- $b$  une fonction déterministe du temps qui permet de répliquer la courbe de taux initiale
- $r_t$  le taux court
- $a$  le paramètre de vitesse de retour à la moyenne
- $\sigma$  la volatilité
- $W_t$  un mouvement Brownien standard
- $a$  et  $\sigma$  sont deux constantes positives

Par absence d'opportunité d'arbitrage, il doit être possible de répliquer la courbe des taux avec l'égalité suivante :

$$b(t) = \frac{\partial f(0, t)}{\partial t} + af(0, t) + \frac{\sigma^2}{2a} (1 - \exp(-2at))$$

Avec  $f(0, t)$  le taux forward instantané tel que :

$$f(t, T) = -\frac{\partial \ln(P(t, T))}{\partial T}$$

Le taux forward de maturité  $T$  vu à l'instant  $t$ , et  $P(t, T)$  le prix en  $t$  d'un zéro-coupon de maturité  $T$ .

En appliquant le lemme d'Itô à  $\exp(at)r_t$ , on obtient :

$$d(\exp(at)r_t) = \exp(at)(b(t)dt + \sigma dW_t)$$

Ainsi pour  $s \leq t$ :

$$r_t = r(s) \exp(-a(t-s)) + \int_s^t \exp(-a(t-u)) b(u) du + \int_s^t \exp(-a(t-u)) \sigma dW_u$$

Les taux étant générés dans le GSE avec une méthode de Monte-Carlo, une discrétisation est réalisée avec un pas de temps mensuel noté  $h = \frac{1}{12}$ . Finalement :

$$r_{(t+h)} = r(t) \exp(-ah) + \lambda(t+h) - \lambda(t) \exp(-ah) + \sigma \sqrt{\frac{1 - \exp(-2ah)}{2a}} Z$$

Avec  $Z \sim \mathcal{N}(0,1)$  et  $\lambda(t) = f(0,t) + \frac{\sigma^2}{2a^2} (1 - \exp(-at))^2$

Le modèle pour l'inflation a été calibré sur l'indice CPXTEMUY avec un historique d'avril 2009 à décembre 2020. Les taux utilisés correspondent à l'Euribor 3 mois sur le même historique. Les paramètres  $a$  et  $\sigma$  ont été calibrés en minimisant l'écart entre les prix théoriques issus de Hull & White et les prix observés sur le marché. Les prix sont obtenus grâce à la courbe des taux et à la volatilité implicite des *swaptions*. La fonction suivante a permis d'effectuer ce calibrage :

$$Obj(a, \sigma) = \sum_{m \in M, t \in T} \left( P_{swaption}^{Hull \& White}(0, m, t, a, \sigma) - P_{swaption}^{marché}(0, m, t) \right)^2$$

Avec :

- $P_{swaption}^{Hull \& White}$  le prix des *swaptions* issu de Hull & White
- $P_{swaption}^{marché}$  le prix des *swaptions* calculé à partir des données de marché
- $m$  la maturité
- $t$  la maturité résiduelle

### GSE univers risque neutre

Cet univers de projection est un univers théorique préconisé notamment par le régulateur pour le calcul de certains indicateurs sous Solvabilité II, tel que le SCR. Deux hypothèses doivent être respectées :

- L'absence d'opportunité d'arbitrage, tel qu'il est impossible de générer un rendement en  $t = 1$  sans risque, avec un investissement nul en  $t = 0$ .
- Les marchés financiers doivent être complets, c'est-à-dire qu'il doit être possible de répliquer les flux d'un portefeuille à partir d'actifs risqués et de l'actif sans risque.

Le modèle de taux utilisé est le même que pour le GSE monde réel car celui-ci respecte les conditions précédentes. L'action et l'immobilier sont générés différemment car en univers risque neutre, le rendement des actifs correspond au taux sans risque.

## Action

Le modèle Black-Scholes à volatilité locale déterministe est utilisé dans ce contexte afin de répliquer la structure par terme de volatilité implicite des *calls* et *puts*. Sous la probabilité risque neutre, ce modèle suit la dynamique suivante :

$$dS_t = S_t(r_t dt + \sigma_t dW_t)$$

Avec :

- $S_t$  le cours de l'action à l'instant  $t$
- $S_0 = 100$
- $r_t$  le taux sans risque issu de notre modèle Hull & White précédent
- $\sigma_t$  la volatilité fonction du temps
- $W_t$  un mouvement Brownien standard sous la probabilité risque neutre  $\mathbb{Q}$

La résolution de cette équation permet de déduire la discrétisation suivante avec un pas de temps annuel noté  $h = 1$  :

$$S_{(t+h)} = S_t \frac{DF(0, t)}{DF(0, t+h)} \exp\left(-\frac{\sigma_{t+h}^2}{2} h + \sigma_{t+h} \sqrt{h} Z\right)$$

Avec :

- $DF(0, t)$  le déflateur, qui vaut  $\exp\left(-\int_0^T r_s ds\right)$
- $Z \sim \mathcal{N}(0,1)$ .

La calibration du modèle s'effectue à partir de la volatilité implicite des options sur l'Euro Stoxx 50, qui permet de déterminer la volatilité locale par pas de temps.

## Immobilier

Le modèle Black-Scholes à volatilité constante est à nouveau utilisé dans ce contexte. Sous la probabilité risque neutre, ce modèle suit la dynamique suivante :

$$dS_t = S_t(r_t dt + \sigma dW_t)$$

Avec :

- $S_t$  le cours de l'action à l'instant  $t$
- $S_0 = 100$
- $r_t$  le taux sans risque issu de notre modèle Hull & White précédent
- $\sigma$  la volatilité du rendement de l'actif sans risque
- $W_t$  un mouvement Brownien standard sous la probabilité risque neutre  $\mathbb{Q}$

La résolution de cette équation permet de déduire la discrétisation suivante avec un pas de temps annuel noté  $h = 1$  :

$$S_{(t+h)} = S_t \exp\left(\int_t^{t+h} r_s ds - \frac{\sigma^2}{2} h\right) + \sigma \sqrt{h} Z$$

Avec  $Z \sim \mathcal{N}(0,1)$ .

## Conclusion

Dans le cadre des travaux menés, l'univers de projection risque neutre sera utilisé en accord avec la réglementation Solvabilité II pour l'évaluation du SCR. Notre indicateur de performance, la PVFP, sera quant à lui évalué en univers monde réel afin de refléter les rendements réellement espérés par les assureurs. Nous utiliserons de plus la PVFP sous univers risque neutre afin d'analyser la performance de prédiction de cet indicateur, et la comparer à celle de la PVFP monde réel.

# Construction de la base de calcul

La méthodologie suivante est mise en place pour la construction de la base de calcul :

- Construction d'un portefeuille d'épargne à partir de diverses hypothèses
- Génération d'une base de données en faisant varier certaines variables dans le modèle ALM
- Vieillessement du portefeuille sur 1 an afin d'évaluer les performances des modèles de *machine learning* après 1 an de vie du portefeuille

Nous verrons dans cette partie la description de chacune de ces étapes, qui permettront par la suite d'utiliser des modèles de *machine learning* afin de calculer la PVFP et le SCR de marché.

## 3.1 Présentation du portefeuille en situation initiale et hypothèses générales

Le portefeuille sur lequel est basée cette étude est un portefeuille d'épargne fictif constitué à dire d'experts et représentatif du marché français. Plusieurs hypothèses générales sont considérées dans le modèle ALM, à savoir :

- le taux concurrentiel (TME) initial défini à -0,34 %, qui correspond au taux OAT 10 ans au 31/12/2020.
- Le taux d'imposition, défini à 34 %
- Le taux de frais d'acquisition de 4 %
- Le taux de frais réel de 0,25 %
- L'inflation à 0 %

La projection sera réalisée pour 1000 simulations sur un horizon de 40 ans. Le bilan en normes comptables françaises (French GAAP) du portefeuille est le suivant :

ACTIF		PASSIF	
	31/12/2020		31/12/2020
<b>ACTIFS INCORPORELS</b>	<b>0</b>	<b>FONDS PROPRES ET RESERVES</b>	<b>51 303 819</b>
		Capitaux propres	0
<b>PLACEMENTS</b>	<b>3 781 824 366</b>	Report à Nouveau	0
Immeubles	279 738 432	Réserve Fonds Garantie	0
Autres placements	3 502 085 934	Réserve de Capitalisation	51 303 819
		Résultat de l'exercice	0
<b>PLACEMENTS DES CONTRATS EN UC</b>	<b>782 721 231</b>	<b>PASSIFS SUBORDONNES</b>	<b>0</b>
<b>CREANCES</b>	<b>0</b>	<b>PROVISIONS TECHNIQUES</b>	<b>3 730 520 547</b>
Autres actifs	0	Provisions d'assurance vie	3 664 558 495
Créances entre sections	0	Provisions pour bénéf. et ristournes	65 962 052
		PRE brute	0
		- dotation PRE restant à constater	0
<b>AUTRES ACTIFS</b>	<b>0</b>	<b>PROVISIONS DES CONTRATS EN UC</b>	<b>782 721 231</b>
Matériel, mobilier	0		
Liquidités	0	<b>PROVISIONS POUR RISQUES ET CHARGES</b>	<b>0</b>
<b>COMPTES DE REGUL-ACTIF</b>	<b>0</b>	<b>DETTES</b>	<b>0</b>
Intérêts courus non échus	0	Dettes diverses	0
Frais d'acquisition reportés	0	Dettes entre sections	0
Frais d'acq/immeubles	0	Dettes à court terme	0
Régl./prix remboursement	0	Dettes fiscales et sociales	0
Régl./primes rmbt emprunts	0		
		<b>COMPTES DE REGUL-PASSIF</b>	<b>0</b>
		Régl./prix remboursement	0
<b>TOTAL GENERAL</b>	<b>4 564 545 597</b>	<b>TOTAL GENERAL</b>	<b>4 564 545 597</b>

**Bilan S1**

Figure 27 - Bilan en norme comptable

L'évolution de la valeur des actifs depuis leur date d'acquisition qui correspond à ce bilan est :

**RECAPITULATIF de l'ETAT des PLACEMENTS 31/12/2020**

RECAPITULATION PAR NATURE		
	Valeur nette au bilan	Valeur de réalisation
Immeubles	279 738 432	311 427 042
Actions et R332-20	325 236 896	382 358 202
OPCVM diversifiés et monétaires		
OPCVM obligataires		
Obligations et R332-19	2 931 673 063	3 091 869 262
Prêts		
Dépôts, et Placements des UC	1 027 897 206	1 027 897 206
Autres actifs représentatifs	-	
<b>ENSEMBLE DES ACTIFS</b>	<b>4 564 545 597</b>	<b>4 813 551 712</b>

Figure 28 - Situation initiale des actifs

Il est noté que la part de cash est de 245 175 975. Les obligations comprennent 52,5 % d'obligations d'entreprises et 47,5 % d'obligations gouvernementales. L'hypothèse des actifs peut être résumée de la façon suivante :

	VNC	Allocation	Plus-value initiale
Actions	325 236 896	8,6%	17,6%
Immobilier	279 738 432	7,4%	11,3%
Cash	245 175 975	6,5%	0%
Obligations	2 931 673 063	77,5%	5,5%

Tableau 1 - Hypothèse des actifs

Avec pour la part d'obligations :

	VNC	Allocation	Plus-value initiale
Obligations d'entreprises	1 538 849 261	52,5 %	3,3 %
Obligations d'état	1 392 823 802	47,5 %	7,9 %

Tableau 2 - Hypothèse des actifs, part d'obligations

Les obligations sont regroupées par maturité résiduelle variant de 1 à 30 ans. Le taux de coupon annuel est recalculé à partir de la VNC et de la valeur de marché pour chaque maturité. Pour les obligations d'entreprises, le taux de coupon en cas du choc de spread prend pour hypothèse une notation de crédit de 2.

Le *turnover* des actions et de l'immobilier est fixé à 10%. Le taux de plus-value initiale du portefeuille est de 6.6%. La maturité pondérée par la VNC des obligations d'entreprises est de 8.8, et celle des obligations gouvernementales de 10.6.

Les caractéristiques principales de notre passif sont les suivantes :

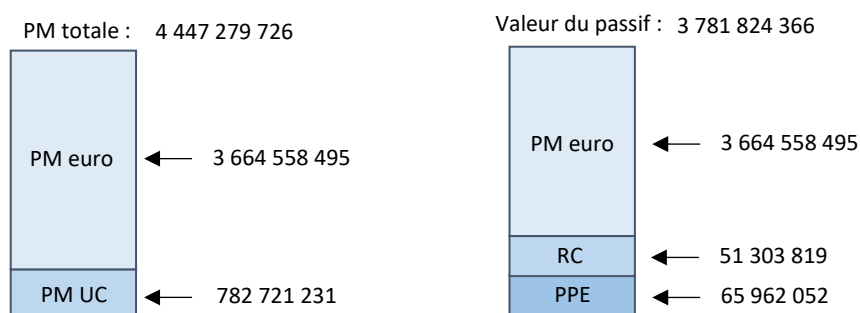


Figure 29 - Caractéristiques du passif

Avec :

- RC : Réserve de Capitalisation
- PPE : Provision pour Participation aux Excédents

L'utilisation du modèle ALM impliquant un temps de calcul relativement conséquent, 94 *model points* regroupant 47 000 polices d'assurance ont été déterminés. En effet, un nombre de *model point* trop élevé rendrait l'évaluation du bilan sous Solvabilité II très chronophage. Différentes variables sont renseignées pour chaque *model point* :

- Le sexe
- L'âge
- L'ancienneté fiscale
- La PM en fonds euros
- La PM en fonds UC
- Le taux de chargements sur encours du fonds euros
- Le taux de chargements sur encours du fonds UC
- Le TMG net de chargements
- Le taux de participation aux bénéfices

La projection dans le cadre Solvabilité II implique une hypothèse de *run-off*, à ce titre, le portefeuille n'a pas de versements libres.

Les hypothèses suivantes sont constantes :

Taux de chargement UC	0,9 %
Taux de chargement euro	0,6 %
TMG	0 %

Tableau 3 - Hypothèses constantes des model points

Un taux de rachats structurels est déterminé en fonction de l'ancienneté fiscale, afin de refléter l'impact de l'avantage fiscal obtenu après 8 ans de détention dans l'assurance-vie. Ainsi, 10% de rachats sont considérés pour une ancienneté fiscale de 5 ans. Les contrats restants sont rachetés progressivement à partir de 8 ans d'ancienneté fiscale.

En pondérant l'ancienneté et l'âge dans nos *model points* par la PM, les caractéristiques suivantes sont obtenues :

Ancienneté moyenne pondérée totale	14,4
Ancienneté moyenne pondérée UC	13,4
Ancienneté moyenne pondérée euro	14,6
Age moyen pondéré total	67,2
Age moyen pondéré UC	66
Age moyen pondéré euro	67,5

Tableau 4 - Caractéristiques des model points

A partir de ces hypothèses, une évaluation du bilan sous Solvabilité II a été réalisée à l'aide du modèle ALM et des tables issues du GSE présenté. Les chocs de marché ont été appliqués et le bilan de la partie correspondant au fond euro récupéré est le suivant :

Bilan S2 (Euro) - Scénario Central			
Actif		Passif	
Obligations	3 480 068 654	PVFP	237 933 260
<i>Etat</i>	<i>1 715 202 813</i>	BEL	4 282 255 609
<i>Corporate</i>	<i>1 764 865 841</i>	Ecart de convergence	(2 215 915)
Actions	382 358 202		
<i>Type 1</i>	<i>382 358 202</i>		
<i>Type 2</i>	-		
Immo	311 427 042		
Cash	344 119 055		
<b>Total</b>	<b>4 517 972 953</b>	<b>Total</b>	<b>4 517 972 953</b>
		<b>EC Relatif</b>	<b>-0,05%</b>

Figure 30 - Bilan sous Solvabilité II



Les SCR correspondant à cette projection sont les suivants :

SCR action	50 969 208
SCR immobilier	23 295 016
SCR hausse de taux	5 506 217
SCR baisse de taux	4 959 679
SCR <i>spread</i>	42 878 292
SCR marché	107 980 572

Tableau 5 - Décomposition du SCR de marché pour le scénario initial

Une projection en monde réel a aussi été réalisée afin d'obtenir la PVFP risque réel et les ratios PVFP/SCR marché.

PVFP monde réel	426 974 279
PVFP risque neutre	237 933 260
Ratio : PVFP monde réel / SCR marché	3,95
Ratio : PVFP risque neutre / SCR marché	2,2

Tableau 6 - Indicateurs principaux pour le scénario initial

## 3.2 Génération de la base de données

A partir de ce scénario initial, l'objectif est de déterminer le couple risque-rendement ici défini par le SCR de marché et la PVFP monde réel pour différentes allocations cible d'actifs. Une base de données a donc été générée à l'aide de l'outil ALM afin d'obtenir le SCR de marché, ainsi que les PVFP en monde réel et risque neutre en modifiant certaines hypothèses. Il a été choisi de faire varier par palier les allocations action, immobilier, cash, obligation ainsi que les plus-values initiales de la part action et immobilier, et le taux de PPE. Les valeurs possibles pour chacun des paramètres sont les suivantes :

Taux cible action	4,5%   9,5%   14,5%   19,5%
Taux cible immobilier	2,7%   7,7%   12,7%
Taux cible cash	1%   6%   11%
Taux cible obligations	100% - Taux cible action + immobilier + cash
Taux plus-values latente actions	12,9%   17,6%   22,3%
Taux plus-values latente immobilier	6,9%   11,3%   15,8%
Ratio PPE / PM	3%   4,5%   6%

Tableau 7 - Valeurs possibles pour chaque paramètre considéré

972 scénarios ont été testés et ont permis de générer une base de données sous la forme suivante :

Scenario	Action	Immo	Cash	Obligs	PVL_action	PVL_immo	PPE	SCR	PVFP_RN	PVFP_RR
0	9,47%	7,71%	6,07%	76,74%	17,56%	11,33%	4,50%	107 980 572	237 933 260	426 974 279
1	4,47%	7,71%	6,07%	81,74%	17,56%	11,33%	4,50%	88 644 841	233 314 811	345 126 873
2	9,47%	7,71%	6,07%	76,74%	17,56%	11,33%	4,50%	107 980 572	237 933 260	426 974 279
3	14,47%	7,71%	6,07%	71,74%	17,56%	11,33%	4,50%	127 403 507	228 174 207	521 810 776

Figure 31 - Extrait de la base de données

La valeur de marché des actifs a aussi été récupéré afin de prendre en compte l'effet taille dans les modèles.

Sur la base des différents scénarios produits en faisant varier nos paramètres, le calcul du SCR de marché et de la PVFP RR est réalisé. L'exploitation de cette base de données permet d'aboutir à la représentation graphique suivante :

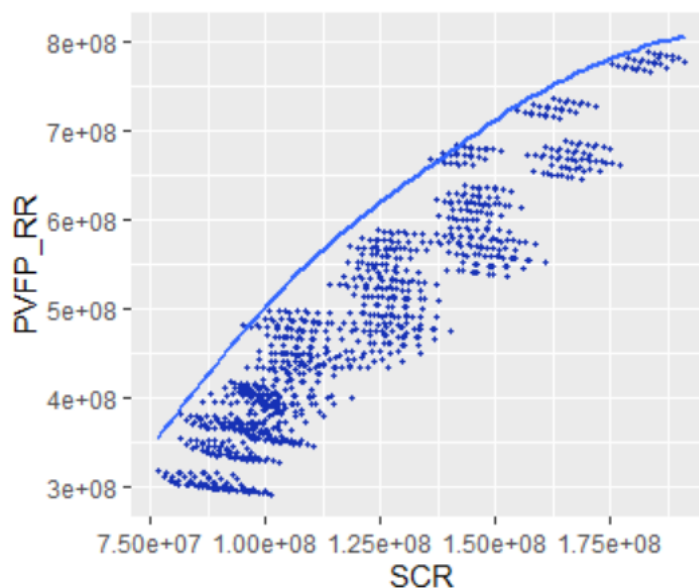


Figure 32 - Représentation du SCR et de la PVFP RR dans la base

La courbe tracée représente la frontière efficiente c'est-à-dire les allocations qui permettent de maximiser la PVFP pour un SCR donné. Le même exercice, désormais réalisé avec le couple SCR marché et PVFP RN permet d'obtenir le graphique suivant :

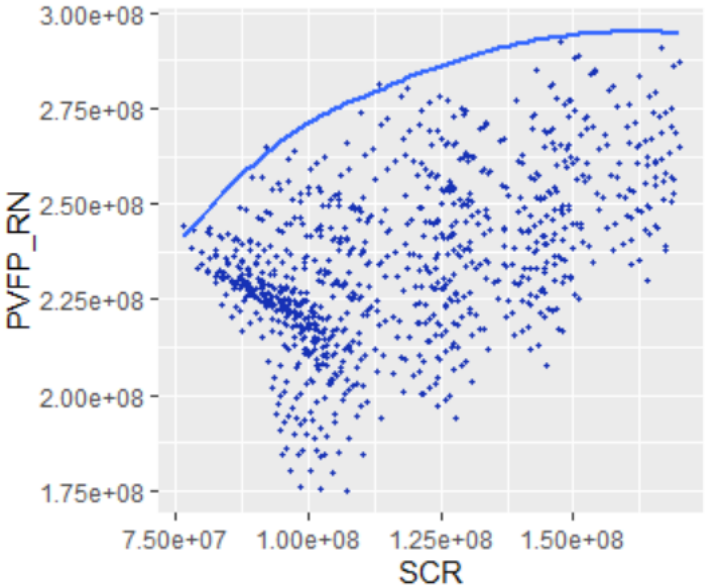
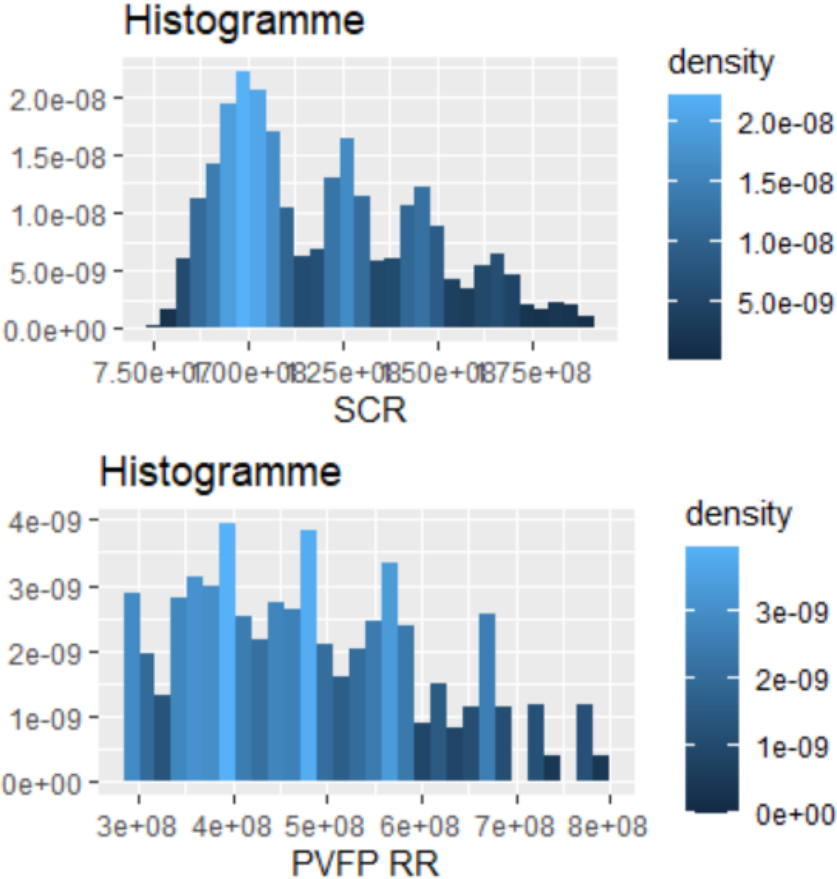


Figure 33 - Représentation du SCR et de la PVFP RN dans la base

Il est possible d'observer les densités correspondantes aux SCR, PVFP risque réel (RR) et PVFP risque neutre (RN) :



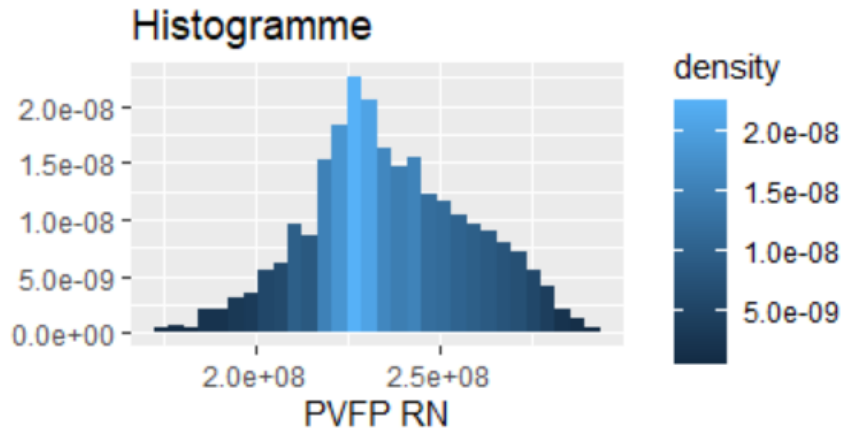


Figure 34 - Densités de la base de données

La densité du SCR ainsi que celle de la PVFP RR sont assez marqués par le fait que la base de données ait été générée en faisant varier les allocations des actifs par pas. Le SCR permet de voir aisément l'impact de cette méthode de génération par pas, les pics de densités plus faibles à droite correspondent à des scénarios où la somme des allocations d'actifs plus risqués (action et immobilier) est grande. Cela représente peu de scénarios car dans le cas du dernier pic à droite, il faut que l'immobilier, l'action et le cash soient à leur allocation maximale.

La matrice suivante met en évidence une corrélation positive entre les allocations d'action, immobilier et cash avec la PVFP risque réel (RR) et le SCR. La PVFP RR et le SCR sont aussi corrélés positivement entre eux. On remarque aussi une corrélation négative entre obligation et les PVFP et SCR. La corrélation négative entre les obligations et les autres actifs est attendue étant donné que l'allocation des obligations est une combinaison linéaire de l'allocation des autres actifs. Il n'y a pas de corrélation entre les plus-values et les PVFP et SCR, néanmoins on observe une légère corrélation négative entre la PPE et la PVFP risque neutre (RN).

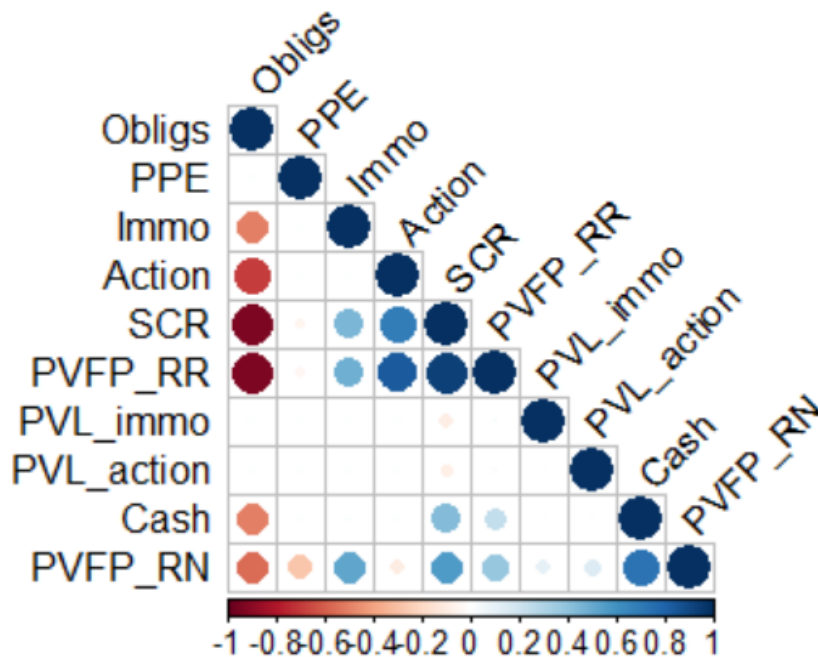


Figure 35 - Matrice des corrélations

La représentation de la PVFP en fonction du SCR sur un repère en mettant en avant la valeur des différentes allocations pour les actifs, et le taux de PPE, est présentée ci-dessous. Pour la PVFP RR les représentations correspondantes sont :

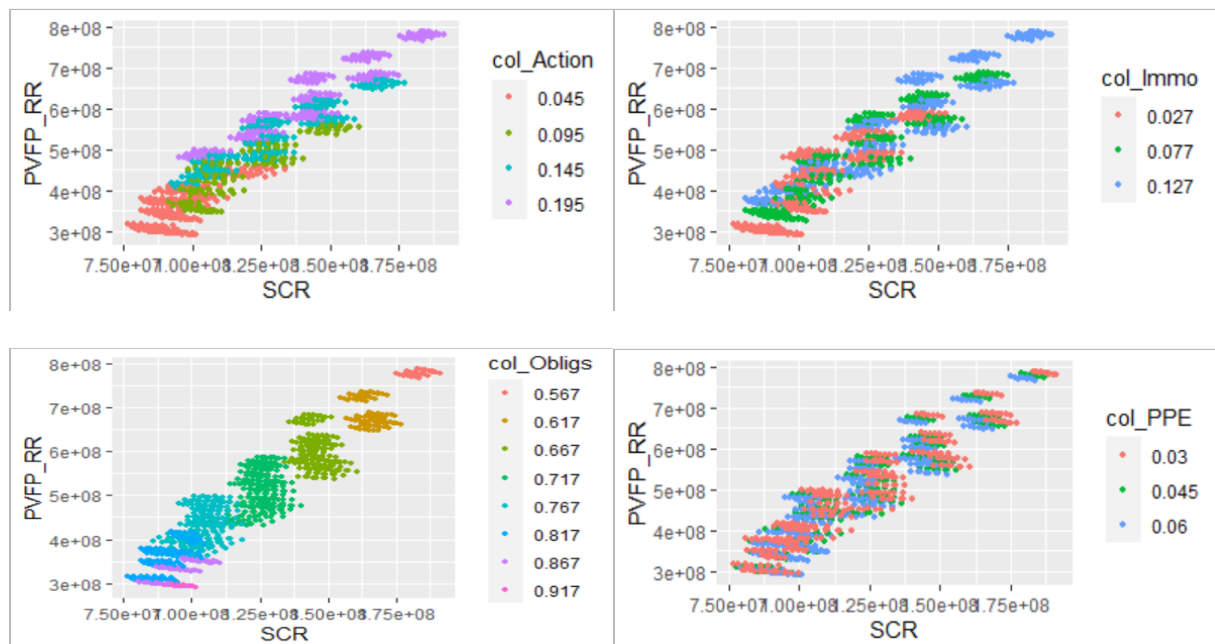


Figure 36 - Représentation des allocations pour la PVFP RR

Pour la PVFP RN, les représentations correspondantes sont :

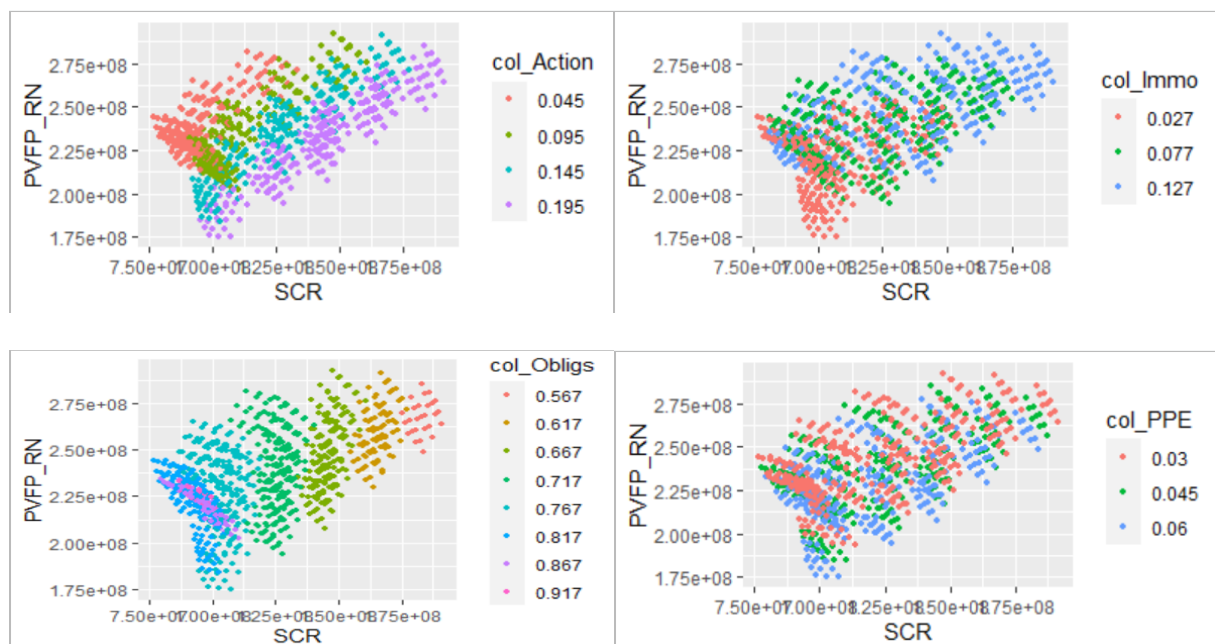


Figure 37 - Représentation des allocations pour la PVFP RN

Avant d'utiliser cette base de données pour construire nos différents modèles de *machine learning*, un vieillissement du portefeuille dans sa situation initiale est effectué. En effet, les modèles seront calibrés sur la base de données présentée, et l'application de ces modèles sur une base issue d'une année de vie du portefeuille permettra de mettre en avant la pertinence de cette méthode pour un assureur.

### 3.3 Vieillessement du portefeuille sur 1 an

Afin de réaliser le vieillissement du portefeuille initial appelé *Portefeuille T<sub>0</sub>*, plusieurs hypothèses ont été considérées. Tout d'abord, il a été choisi de conserver les tables issues du GSE risque neutre en cristallisant les prix et taux des zéro coupon à chaque pas de maturité. Les rendements des actifs action et immobilier correspondent quant à eux aux tables générées en environnement monde réel afin de représenter une évolution de la valeur des actifs cohérente avec les attentes d'un assureur.

A partir de ces hypothèses, le *Portefeuille T<sub>0</sub>* est projeté à partir du scénario central, sur une année. Le modèle ALM permet notamment d'obtenir le montant de l'actif et du passif, mais aussi la table issue de la simulation des *model points* après un an de vie. Les informations issues de ce vieillissement permettent de paramétrer le *Portefeuille T<sub>1</sub>* ;

L'hypothèse des actifs de ce nouveau portefeuille est désormais :

	VNC	Plus-value initiale
Actions	330 432 281	21,5 %
Immobilier	285 750 973	14,4 %
Cash	231 651 746	0 %
Obligations	2 780 138 382	17 %

Tableau 8 - Hypothèse des actifs du portefeuille T<sub>1</sub>

Avec pour la part d'obligations :

	VNC	Plus-value initiale
Obligations d'entreprises	1 506 719 381	21,4 %
Obligations d'état	1 273 418 998	13,3 %

Tableau 9 - Hypothèse des actifs du portefeuille T<sub>1</sub>, part des obligations

Les caractéristiques principales de notre passif sont les suivantes :

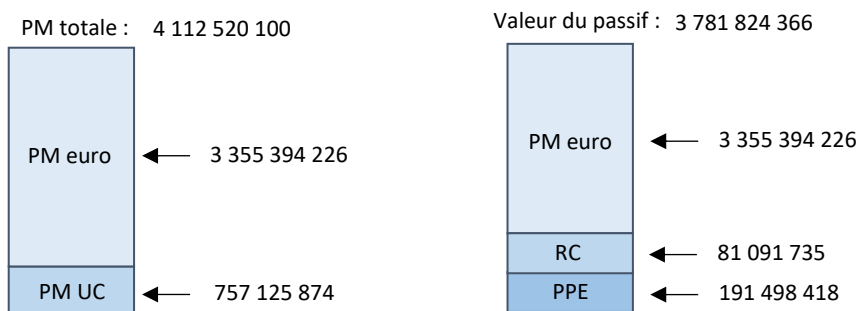


Figure 38 - Caractéristiques du passif du portefeuille T<sub>1</sub>

48 scénarios ont été testés et ont permis de générer une nouvelle base de données au format similaire. Cette dernière permettra de tester l'efficacité des modèles construits sur le *Portefeuille T<sub>0</sub>*, et appliqués sur le *Portefeuille T<sub>1</sub>* afin de déterminer la robustesse des modèles après une évolution d'une année du portefeuille initial.

De la même manière, un second portefeuille en  $T_1$  intégrant de nouveaux contrats a été généré. En accord avec une observation d'experts de portefeuilles représentatifs du marché, une hypothèse de 10% de collecte brut est prise en compte. Un versement libre de 360 000 000 sur le fonds euros et de 90 000 000 sur le fonds UC avec un chargement sur prime de 3 % est ainsi intégré au vieillissement. L'hypothèse des actifs de ce second portefeuille est désormais :

	VNC	Plus-value initiale
Actions	364 528 545	19,5 %
Immobilier	313 522 049	13,2 %
Cash	243 207 783	0 %
Obligations	3 029 023 861	16,5 %

Tableau 10 - Hypothèse des actifs du portefeuille  $T_1$  avec nouveaux contrats

Avec pour la part d'obligations :

	VNC	Plus-value initiale
Obligations d'entreprises	1 647 453 633	21,4 %
Obligations d'état	1 381 570 229	12,4 %

Tableau 11 - Hypothèse des actifs du portefeuille  $T_1$  avec nouveaux contrats, part des obligations

Les caractéristiques principales de notre passif sont les suivantes :

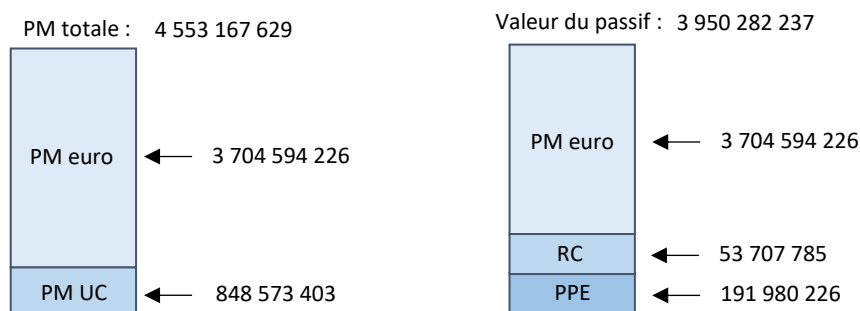


Figure 39 - Caractéristiques du passif du portefeuille  $T_1$  avec nouveaux contrats

Les bases de données sont maintenant générées et exploitables. La prochaine section de ce mémoire présentera les différents modèles utilisés sur ces données, et visera à répondre à notre problématique.

# Utilisation du *machine learning* pour déterminer l'allocation optimale

Plusieurs modèles ont été considérés dans le cadre de ces travaux. Leur construction est présentée ci-après. Dans un premier temps, le choix des paramètres sélectionnés ainsi que celui de leur calibration seront présentés. Une étude de la performance des résultats sur le portefeuille  $T_0$  sera ensuite réalisée, et l'application des modèles en  $T_1$  permettra d'évaluer la pertinence des différentes méthodes. Enfin, une étude de la sensibilité des modèles par rapport à la variation de la valeur de marché de l'action et de l'immobilier permettra de juger leur robustesse sous un angle différent.

## 4.1 GLM

A partir de la base de données, les premiers modèles réalisés sont des modèles GLM car ils sont très souvent utilisés dans le monde de l'assurance en raison de leur explicabilité. L'objectif est de pouvoir déterminer l'allocation cible optimale au regard du taux de PMVL et du ratio de PPB/PM permettant de maximiser la PVFP RR sous une contrainte de niveau de SCR de marché. D'autres modèles ont également été développés afin d'évaluer la performance de prédiction sur la PVFP RN et les ratios PVFP/SCR.

### Choix des paramètres

2 composantes sont à déterminer pour chaque modèle :

- Le choix de la loi
- Le choix des variables prises en compte

La base de données  $T_0$  a été séparée en deux, une base d'apprentissage qui permet de calibrer les coefficients  $\beta$  et de construire le modèle, et une base de test qui permet d'apprécier l'efficacité du modèle sur un échantillon de données qu'il ne connaît pas. La base d'apprentissage représente 80% des individus de notre base de données, après tirage aléatoire, et la base de test représente les 20% restants.

### **Choix de la loi**

La sélection de ce paramètre est issue de l'analyse de la densité des différents indicateurs et de leur fonction de répartition. Il est important de noter que l'analyse de la densité est délicate car la construction de la base par palier est un élément qui se retrouve sur les histogrammes. On peut observer par exemple pour le SCR que chaque morceau de la densité correspond à des combinaisons d'allocation d'actifs différentes. La partie à droite correspond aux scénarios avec une combinaison d'allocation actions et immobilier maximale, et les parties plus à gauche sont constituées de scénarios avec des allocations actions et immobilier plus faibles.

La densité des indicateurs est donc un élément à considérer avec précaution. La fonction de répartition permettra néanmoins d'effectuer une analyse plus pertinente. A l'issue de notre analyse, la loi log-normale est retenue. Les densités et fonctions de répartition des différents indicateurs en fonction de la loi log-normale sont les suivantes :



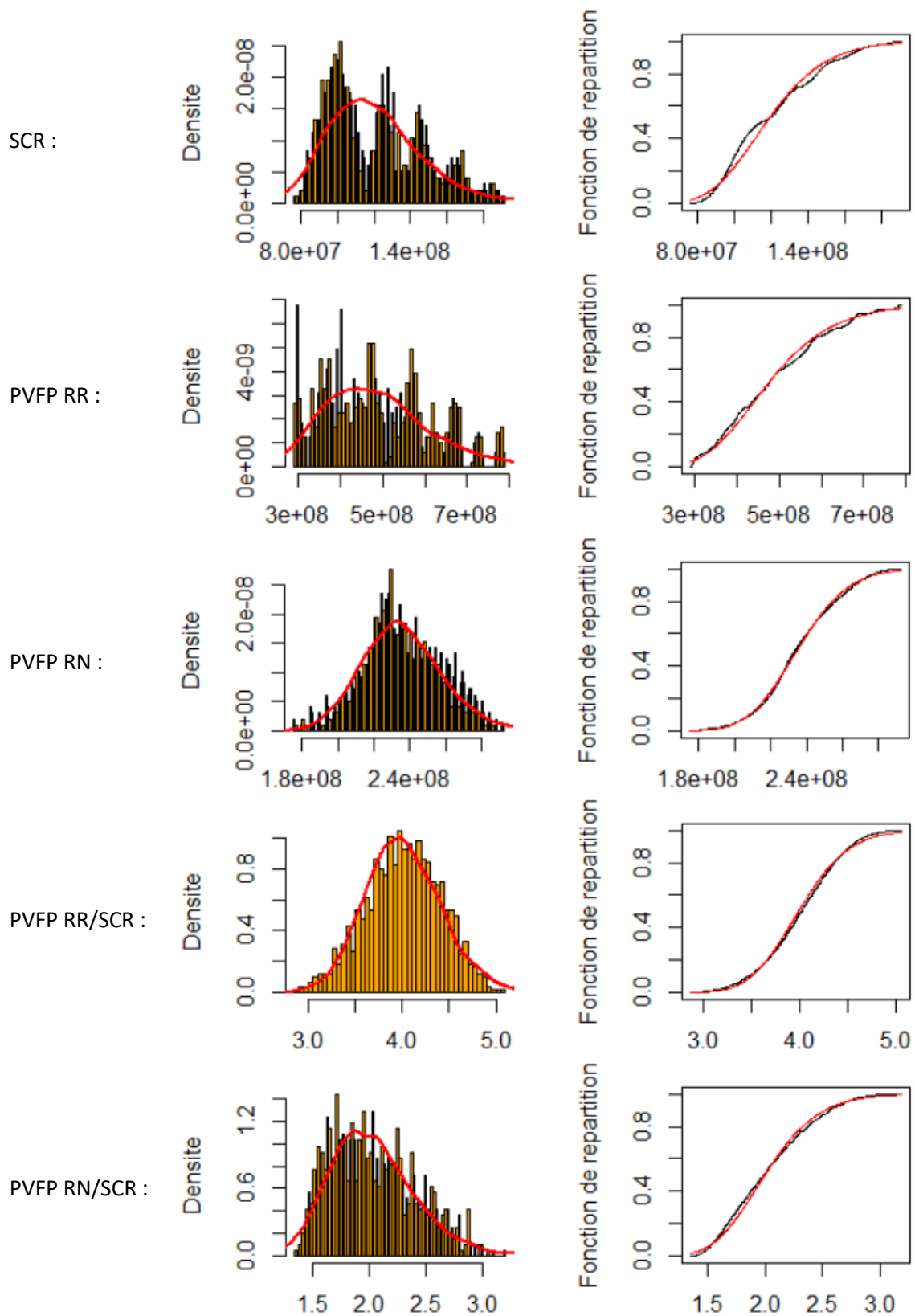


Figure 40 - GLM : Représentation de la densité et des fonctions de répartition pour chaque indicateur

Afin de confirmer que chacun des indicateurs suit bien la loi log-normale, un test de Kolmogorov-Smirnov a été réalisé. Pour un seuil d'acceptation  $\alpha = 5\%$ , la *p-value* confirme l'hypothèse de loi log-normale.

## Sélection des variables

La méthode de sélection des variables retenue est la méthode descendante. Pour rappel, elle consiste à implémenter le modèle tout d'abord avec l'ensemble des variables. La variable la moins significative est ensuite enlevée du modèle jusqu'à un certain seuil. La significativité des variables est définie par son effet sur le critère d'information d'Akaike noté AIC :

$$AIC = -2L + 2k$$

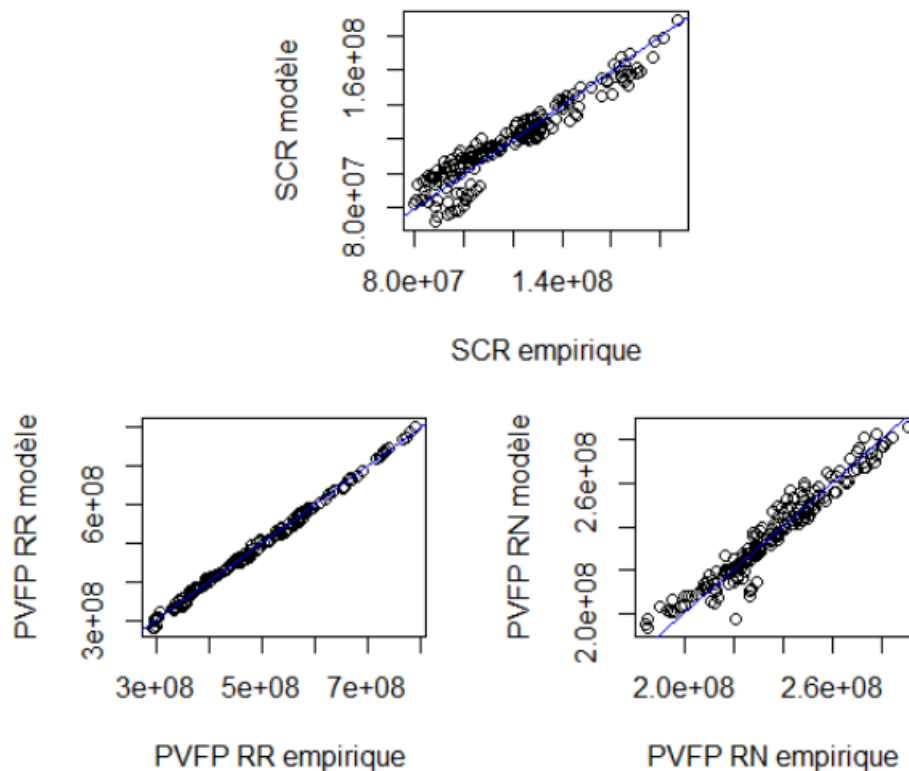
Avec  $L$  la log-vraisemblance du modèle estimé et  $k$  le nombre de paramètres du modèle.

Le résultat est similaire pour chacun des indicateurs, le taux d'allocation des obligations n'est pas pris en compte car c'est une combinaison linéaire des autres taux d'allocation. Il en est de même pour les taux de plus-values latentes actions et immobilier avec leurs valeurs de marché. Enfin, la PPE et la valeur de marché de la part de *cash* sont fortement corrélées, ainsi le modèle ne retiendra que la PPE. Les modèles retiendront ainsi les variables suivantes :

- Taux cible action
- Taux cible immobilier
- Taux cible *cash*
- Valeur de marché des actions
- Valeur de marché de l'immobilier
- Ratio PPE/PM

## Résultats $T_0$

Les modèles, créés sur la base d'apprentissage avec les paramètres précédemment déterminés, ont ensuite été appliqués sur la base de test. La représentation des valeurs théoriques et empiriques des différents indicateurs est la suivante :



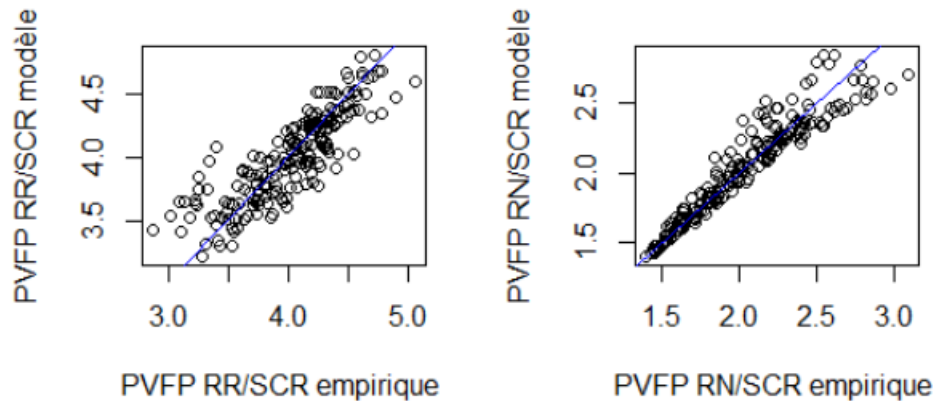


Figure 41 - GLM : représentation des indicateurs théoriques et calculés

Les qualités de prédiction sont assez variables en fonction des modèles. Parmi les différentes métriques d'erreur généralement utilisées, nous considérons RMSE car :

- Le MAE et MAPE ne pénalisent pas suffisamment les écarts importants et la base ne comprend pas de valeurs extrêmes
- Le RMSE permet de ramener l'erreur du MSE à son unité de base en appliquant la racine carrée

Nous calculons l'erreur moyenne telle que donnée par l'expression :

$$\text{Erreur moyenne} = \frac{RMSE}{\bar{Y}}$$

Avec  $\bar{Y}$  la moyenne de l'indicateur de la base de test et :

$$RMSE = \sqrt{E[(\hat{Y} - Y)^2]}$$

Avec  $\hat{Y}$  le vecteur d'estimation de l'indicateur, et  $Y$  le vecteur de l'indicateur de la base de test.

Une étude des résultats sur 1 000 modèles permet d'obtenir les statistiques suivantes :

Modèles	Erreur moyenne	min	max
SCR	6,17 %	5,36 %	7,09 %
PVFP RR	1,53 %	1,32 %	1,77 %
PVFP RN	2,61 %	2,13 %	3,17 %
PVFP RR/SCR	5,32 %	4,53 %	6,23 %
PVFP RN/SCR	5,48 %	4,52 %	6,46 %

Tableau 12 - Performance de la GLM en  $T_0$

Les modèles de prédiction des PVFP sont intéressants, les estimations relatives au SCR ont une erreur moyenne qui est néanmoins plutôt élevée. L'étude des résidus est semblable pour tous les modèles. Dans le cas de la PVFP RR, la représentation des résidus de Pearson montre les points du nuage uniformément répartis et sans structure apparente. L'hétéroscédasticité est donc bien corrigée :

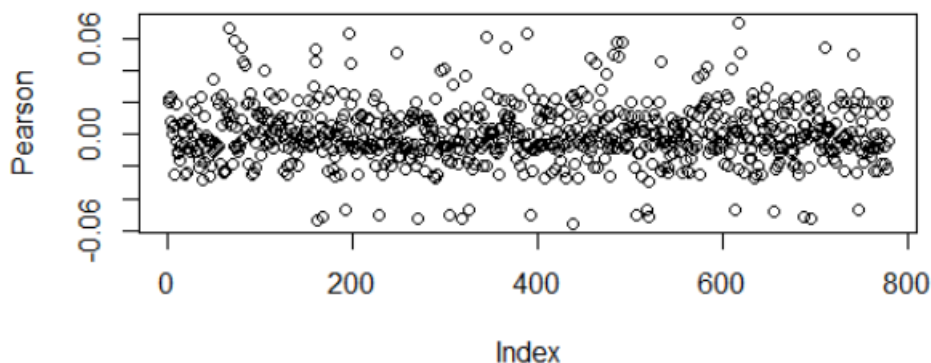


Figure 42 - GLM : représentation des résidus de Pearson pour la PVFP RR

### Résultats $T_1$

Les modèles ont été testés sur la base issue du *Portefeuille  $T_1$*  sans intégration de nouveaux contrats. Les résultats obtenus sur 1 000 modèles sont les suivants :

Modèles	Erreur moyenne	min	max
SCR	7,98 %	7,45 %	8,67 %
PVFP RR	5,10 %	4,86 %	5,37 %
PVFP RN	4,85 %	4,55 %	5,15 %
PVFP RR/SCR	10,78 %	10,13 %	11,54 %
PVFP RN/SCR	10,83 %	10,17 %	11,40 %

Tableau 13 - Performance de la GLM en  $T_1$

Les résultats sont corrects pour les PVFP après le vieillissement, néanmoins l'erreur moyenne des modèles SCR et ratios PVFP/SCR est élevée, ce qui rend peu adaptée l'utilisation d'une GLM dans le contexte de nos travaux. Le même exercice est réalisé avec la base de données issue du *Portefeuille  $T_1$*  avec intégration de nouveaux contrats. Les statistiques des modèles sur cette nouvelle base sont :

Modèles	Erreur moyenne	min	max
SCR	26,23 %	24,66 %	27,75 %
PVFP RR	11,92 %	11,21 %	12,63 %
PVFP RN	26,82 %	25,52 %	28,03 %
PVFP RR/SCR	48,82 %	46,36 %	51,52 %
PVFP RN/SCR	69,89 %	67,21 %	72,24 %

Tableau 14 - Performance de la GLM en  $T_1$  avec nouveaux contrats

Les modèles ne sont pas exploitables dans ce cas. Dans le cas du SCR, la moyenne de notre base de données est de 140,9 millions, et le modèle sous-estime exclusivement son estimation. Le résultat obtenu sera donc compris entre 101,8 millions et 106,1 millions.

## Test de sensibilité aux plus-values

Un test de sensibilité est réalisé sur le scénario central du *Portefeuille T<sub>1</sub>* sans nouveaux contrats pour le calcul de la PVFP RR. Les variables considérées sont les suivantes :

Taux cible action	9,5%
Taux cible immobilier	7,7%
Taux cible <i>cash</i>	6%
Taux cible obligations	76,8 %
Taux plus-values latente actions	21,5 %
Taux plus-values latente immobilier	14,4 %
Ratio PPE / PM	5,7 %
VM action	401 475 222
VM immobilier	326 899 113
VM <i>cash</i>	231 651 746
VM obligations	3 252 916 481

Tableau 15 - Caractéristiques du scénario initial après vieillissement du portefeuille

1 000 modèles ont été lancés sur ce scénario en situation de hausse ou de baisse des plus-values actions et immobilier. Les valeurs de marchés ont été recalculées avec une nouvelle hypothèse de taux de plus-value latentes. Les situations suivantes ont été testées afin d'obtenir des statistiques sur la PVFP RR calculée :

Situation	Moyenne	min	max
-75 % PMVL action et immobilier	379 333 221	366 778 570	389 919 810
-50 % PMVL action et immobilier	390 946 103	381 729 711	398 802 235
-25 % PMVL action et immobilier	402 917 924	397 290 311	407 887 003
0 % PMVL action et immobilier	415 259 884	413 361 640	417 178 723
25 % PMVL action et immobilier	427 983 534	425 853 266	430 340 276
50 % PMVL action et immobilier	441 100 788	435 399 487	447 882 408
75 % PMVL action et immobilier	454 623 937	445 133 615	466 139 617
100 % PMVL action et immobilier	468 565 659	455 085 367	485 141 052
125 % PMVL action et immobilier	482 939 030	465 259 608	504 917 049
150 % PMVL action et immobilier	497 757 540	475 661 312	525 499 184
175 % PMVL action et immobilier	513 035 105	486 295 565	546 920 316
200 % PMVL action et immobilier	528 786 078	497 167 565	569 214 647

Tableau 16 - Test de sensibilité aux PMVL de la GLM

La variation des plus-values est bien prise en compte dans les résultats du modèle. L'effet de l'augmentation des PMVL actions et immobiliers est linéaire, et peut être observé sur le graphique suivant :

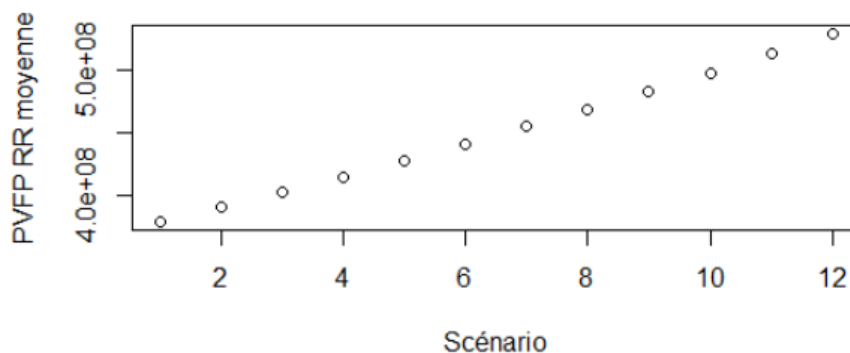


Figure 43 - GLM : sensibilité aux PMVL

## 4.2 Random Forest

A partir de la base de données, des modèles *Random Forest* ont été réalisés afin de pouvoir prédire le SCR de marché et la PVFP RR associés en faisant varier certains paramètres prédéfinis. L'objectif est de pouvoir déterminer l'allocation cible optimale au regard du taux de PMVL et du ratio de PPB/PM permettant de maximiser la PVFP RR sous une contrainte de niveau de SCR de marché. D'autres modèles ont également été développés afin d'évaluer la performance de prédiction sur la PVFP RN et les ratios PVFP/SCR.

### Choix des paramètres

3 hyper-paramètres sont considérés pour la construction de la forêt :

- `ntree` qui correspond au nombre d'arbres
- `mtry` qui correspond au nombre de variables à utiliser dans la construction de chaque arbre
- `maxnodes` qui correspond au nombre maximal de nœuds par arbre

La base de données  $T_0$  a été séparée en deux, une base d'apprentissage qui permet de calibrer les paramètres et de construire le modèle, et une base de test qui permet d'apprécier l'efficacité du modèle sur un échantillon de données qu'il ne connaît pas. Comme dans le modèle précédent, la base d'apprentissage représente 80% des individus de notre base de données, après tirage aléatoire, et la base de test représente les 20% restants.

### **ntree**

Sur la base d'apprentissage, la sélection de `ntree` a été réalisée en traçant le MSE de la forêt en fonction du nombre d'arbres avec les paramètres par défaut. Le graphique est similaire pour les deux PVFP et le SCR et ressemble au graphique suivant :

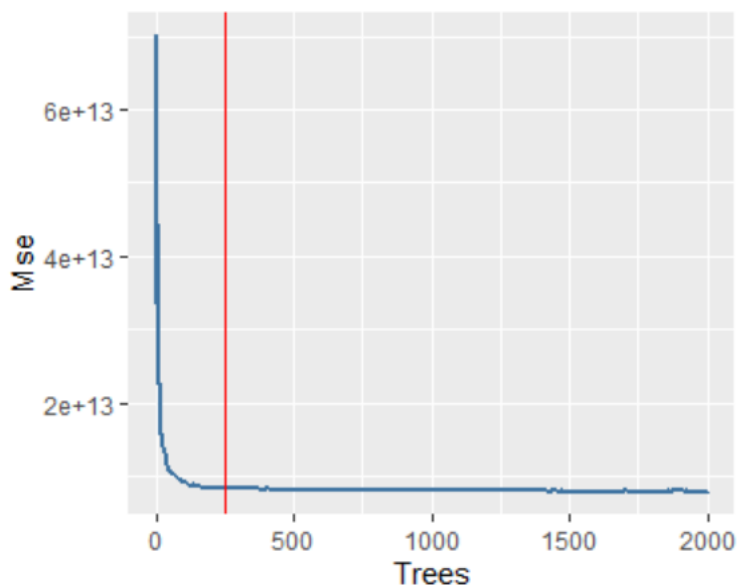


Figure 44 - Random Forest : sélection du paramètre `ntree` en fonction du MSE

Cela permet de définir que l'erreur se stabilise à partir de 250 arbres en général et donc de fixer ce paramètre à 250.

## mtry et maxnodes

Pour le choix de mtry et de maxnodes, le RMSE a été calculé en faisant varier mtry de 2 à 11 (nombre de variables explicatives) et maxnodes de 1 à 100 par pas de 1. Le graphique représentant le RMSE moyen par nombre de variables à utiliser pour le calcul du SCR est le suivant :

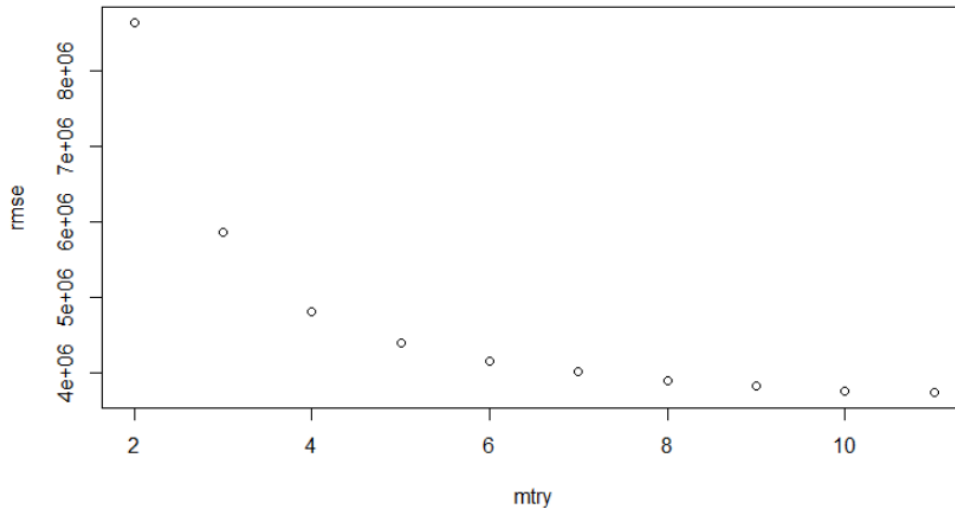


Figure 45 - Random Forest : sélection du paramètre mtry en fonction du RMSE pour le SCR

Afin d'éviter un phénomène de sur-apprentissage, le paramètre recherché doit être le plus faible possible, tout en ayant un RMSE faible. Le graphique ne permet pas d'obtenir un point d'inflexion précis, néanmoins nous considérerons une valeur de mtry = 6 pour le SCR. Le RMSE moyen par nombre maximal de nœuds par arbre à utiliser pour le calcul du SCR peut être représenté de la façon suivante :

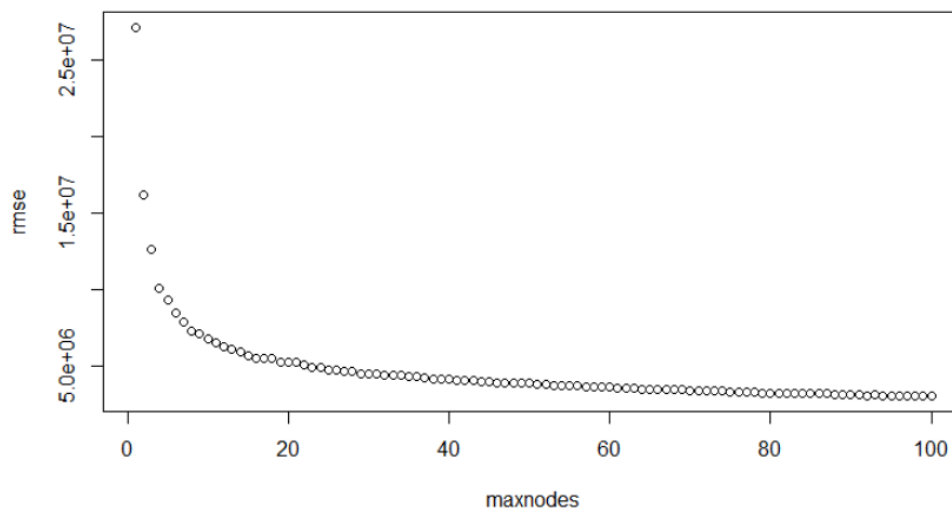


Figure 46 - Random Forest : sélection du paramètre maxnodes en fonction du RMSE pour le SCR

Le paramètre maxnodes retenu dans la construction de ce modèle sera 30. En appliquant la même démarche sur la PVFP RR, le graphique suivant est obtenu :

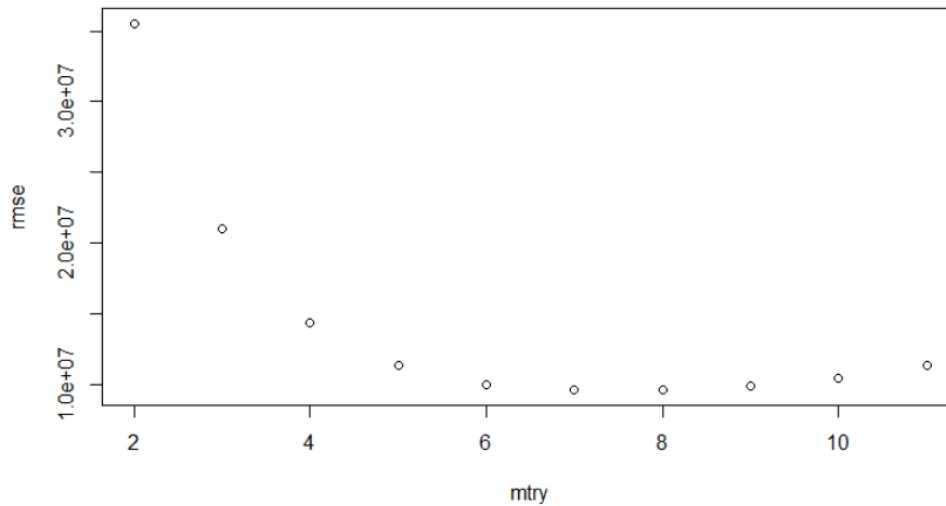


Figure 47 - Random Forest : sélection du paramètre mtry en fonction du RMSE pour la PVFP RR

Cette fois, l'identification du point d'inflexion est plus flagrante. Le mtry retenu sera une fois encore de 6. L'étude du choix de ce paramètre pour la PVFP RN donne des résultats similaires. Ainsi, la même valeur de mtry est retenue. Le même exercice est réalisé pour les ratios et permet de définir cette fois mtry = 5 :

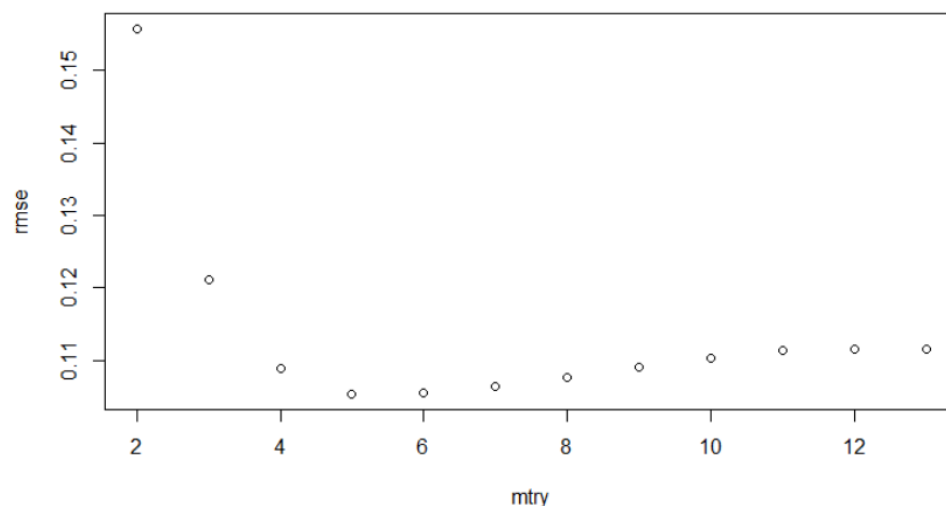


Figure 48 - Random Forest : sélection du paramètre mtry en fonction du RMSE pour le ratio PVFP RR/SCR

L'analyse du paramètre maxnodes dans le cas des PVFP et des ratios SCR/PVFP donne un graphique semblable à celui du SCR, le paramètre de 30 sera aussi retenu. Les paramètres retenus pour les différents modèles sont donc :

Modèles	ntree	maxnodes	mtry
SCR	250	30	6
PVFP RR	250	30	6
PVFP RN	250	30	6
PVFP RR/SCR	250	30	5
PVFP RN/SCR	250	30	5

Tableau 17 - Paramétrage des modèles random forest



## Résultats T<sub>0</sub>

Les modèles, créés sur la base d'apprentissage avec les paramètres précédemment déterminés, ont ensuite été appliqués sur la base de test. La représentation des valeurs théoriques et empiriques des différents indicateurs est la suivante :

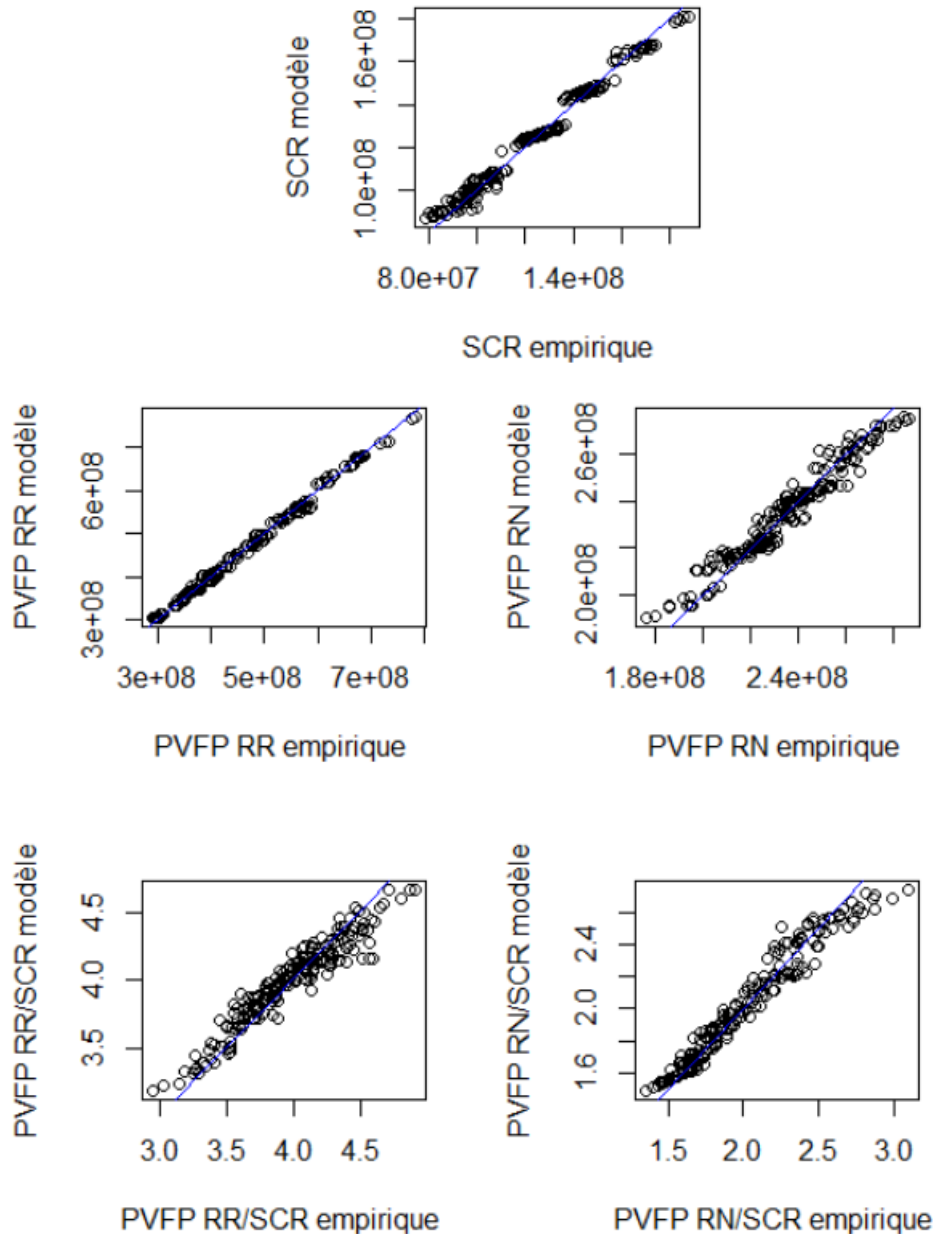


Figure 49 - Random Forest : représentation des indicateurs théoriques et calculés

Les qualités de prédiction sont à priori assez intéressantes. L'erreur moyenne définie pour la GLM est à nouveau utilisée, telle que donnée par l'expression :

$$\text{Erreur moyenne} = \frac{RMSE}{\bar{Y}}$$

Avec  $\bar{Y}$  la moyenne de l'indicateur de la base de test. Ce taux est faible dans chacun des cas. Une étude des résultats sur 1 000 modèles permet d'obtenir les statistiques suivantes :

Modèles	Erreur moyenne	min	max
SCR	2,97 %	2,53 %	3,50 %
PVFP RR	1,84 %	1,46 %	2,19 %
PVFP RN	2,26 %	1,92 %	2,68 %
PVFP RR/SCR	2,89 %	2,35 %	3,39 %
PVFP RN/SCR	4,21 %	3,12 %	5,37 %

Tableau 18 - Performance de la random forest en  $T_0$

La forêt aléatoire permet aussi de récupérer l'importance des variables utilisées dans sa construction. L'ordre d'importance des variables est similaire pour les modèles SCR, PVFP RR et leur ratio :

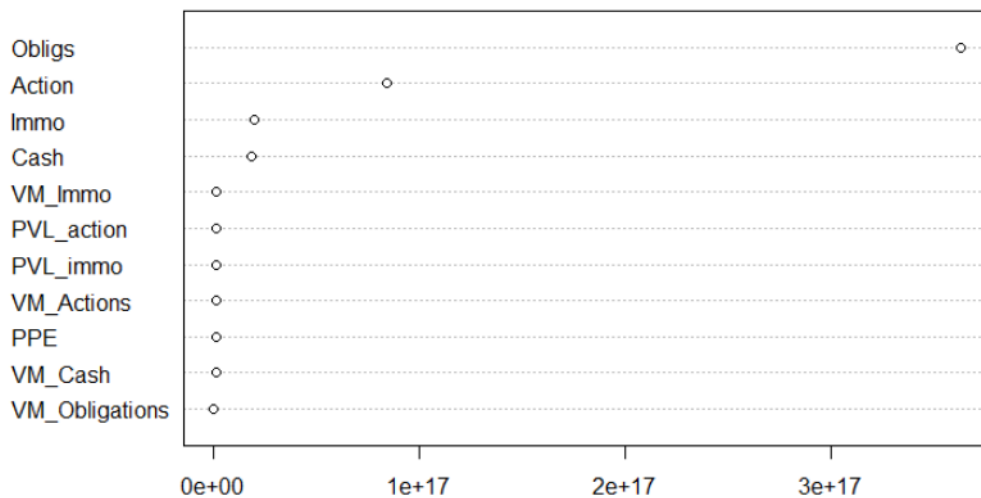


Figure 50 - Random Forest : ordre d'importance des variables pour le SCR

Pour la prédiction du SCR et de la PVFP RR, il est observé que l'obligation est la variable qui a la plus grande importance, ce qui s'explique par le fait qu'elle est la combinaison linéaire des autres allocations d'actifs, or l'allocation des actions et de l'immobilier a un grand impact sur la PVFP RR et le SCR.

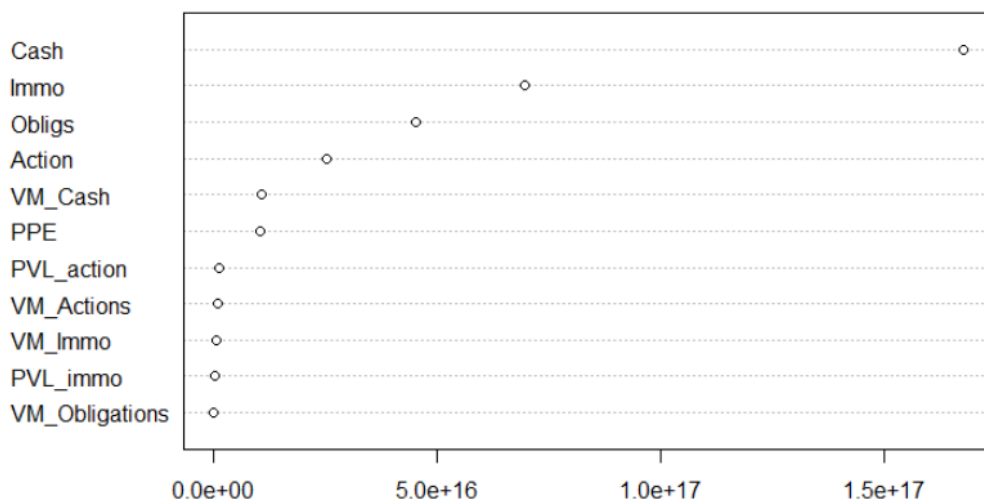


Figure 51 - Random Forest : ordre d'importance des variables pour la PVFP RN

Cependant, pour la PVFP RN et le ratio, l'importance des variables change. Toutes les variables ont désormais un impact relativement important sauf les taux de plus-value et les VM. Cela s'explique

par les rendements des actifs qui sont désormais similaires au taux sans risque en environnement risque neutre.

Pour tous les modèles, l'importance des VM est très limitée car elle varie trop peu dans le portefeuille pour avoir un impact lors de la création des arbres. Seule la valeur de marché de la part de *cash* montre une certaine importance, car son montant est celui qui varie le plus parmi les VM.

### Résultats $T_1$

Les modèles ont été testés sur la base issue du *Portefeuille  $T_1$*  sans intégration de nouveaux contrats. Les résultats obtenus sur 1 000 modèles sont les suivants :

Modèles	Erreur moyenne	min	max
SCR	3,39 %	3,15 %	3,64 %
PVFP RR	4,77 %	4,59 %	5,00 %
PVFP RN	3,08 %	2,81 %	3,39 %
PVFP RR/SCR	4,65 %	4,23 %	5,17 %
PVFP RN/SCR	6,62 %	5,76 %	7,32 %

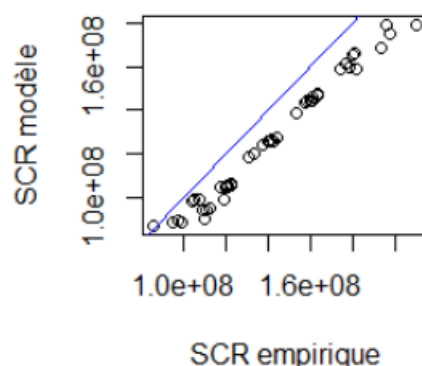
Tableau 19 - Performance de la random forest en  $T_1$

Les résultats restent très bons après le vieillissement, l'erreur moyenne du modèle PVFP RR est celle qui augmente le plus en passant de 1,8 % à 4,8 %. Le même exercice est réalisé avec la base de données issue du *Portefeuille  $T_1$*  avec intégration de nouveaux contrats. Les statistiques des modèles sur cette nouvelle base sont :

Modèles	Erreur moyenne	min	max
SCR	10,83379 %	10,47 %	11,23 %
PVFP RR	8,81107 %	8,63 %	8,98 %
PVFP RN	16,68325 %	16,38 %	17,00 %
PVFP RR/SCR	22,42147 %	21,99 %	22,83 %
PVFP RN/SCR	33,65730 %	32,92 %	34,39 %

Tableau 20 - Performance de la random forest en  $T_1$  après intégration de nouveaux contrats

Les résultats sont beaucoup moins bons. La représentation des valeurs théoriques et empiriques des différents indicateurs permet d'observer un biais important. Les graphiques sont les suivants :



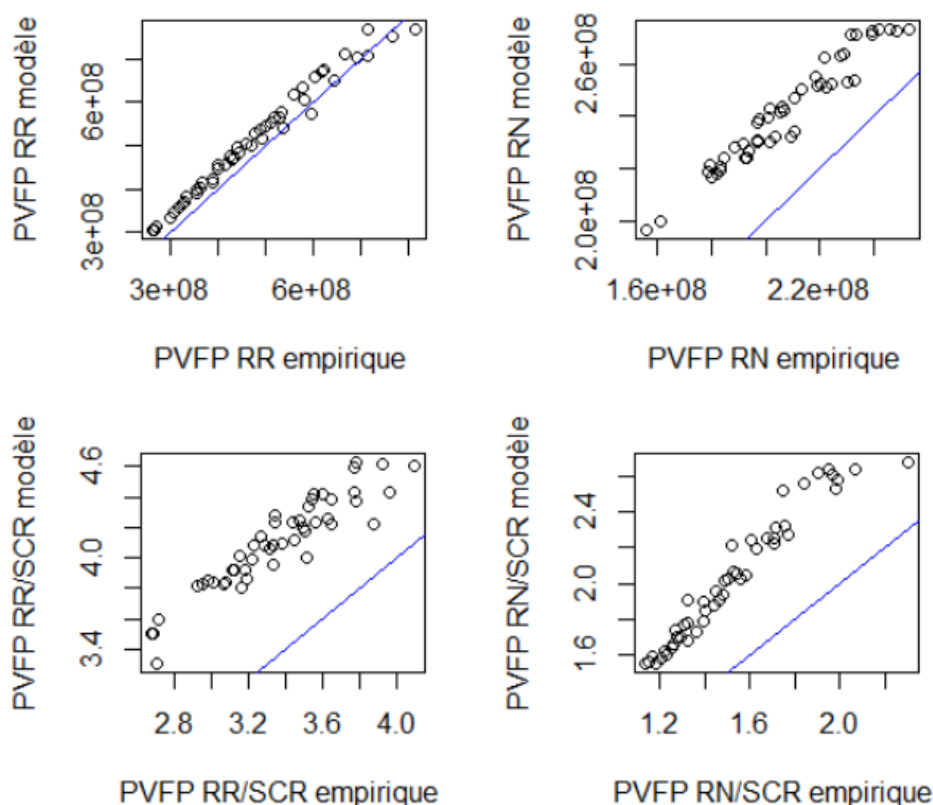


Figure 52 - Random Forest : représentation des indicateurs théoriques et calculés T1 avec nouveaux contrats

Les biais s'expliquent en partie par une forte variation des VM à la suite du vieillissement intégrant de nouveaux contrats. La sous-estimation du SCR est due à la VM des actifs étant globalement plus grande dans ce scénario et les modèles prenant peu en compte son effet. Le SCR calculé par le modèle ALM augmente ainsi, et non celui des *random forest*.

#### Test de sensibilité aux plus-values

Un test de sensibilité est réalisé sur le scénario central du *Portefeuille T<sub>1</sub>* sans nouveaux contrats pour le calcul de la PVFP RR. Les variables considérées sont les suivantes :

Taux cible action	9,5%
Taux cible immobilier	7,7%
Taux cible <i>cash</i>	6%
Taux cible obligations	76,8 %
Taux plus-values latente actions	21,5 %
Taux plus-values latente immobilier	14,4 %
Ratio PPE / PM	5,7 %
VM action	401 475 222
VM immobilier	326 899 113
VM <i>cash</i>	231 651 746
VM obligations	3 252 916 481

Tableau 21 - Caractéristiques du scénario initial après vieillissement du portefeuille

1 000 modèles ont été lancés sur ce scénario en situation de hausse ou de baisse des plus-values actions et immobilier. Les valeurs de marchés ont été recalculées avec une nouvelle hypothèse

de taux de plus-value latentes. Les situations suivantes ont été testées afin d'obtenir des statistiques sur la PVFP RR calculée :

Situation	Moyenne	min	max
-75 % PMVL action et immobilier	422 801 450	416 442 816	428 621 232
-50 % PMVL action et immobilier	422 801 450	416 442 816	428 621 232
-25 % PMVL action et immobilier	422 801 450	416 442 816	428 621 232
0 % PMVL action et immobilier	422 801 450	416 442 816	428 621 232
25 % PMVL action et immobilier	423 077 580	416 809 737	429 935 816
50 % PMVL action et immobilier	423 077 580	416 809 737	429 935 816
75 % PMVL action et immobilier	423 077 580	416 809 737	429 935 816
100 % PMVL action et immobilier	423 077 580	416 809 737	429 935 816
125 % PMVL action et immobilier	423 077 580	416 809 737	429 935 816
150 % PMVL action et immobilier	423 077 580	416 809 737	429 935 816
175 % PMVL action et immobilier	423 077 580	416 809 737	429 935 816
200 % PMVL action et immobilier	423 077 580	416 809 737	429 935 816

*Tableau 22 - Test de sensibilité aux PMVL de la random forest*

La variation des plus-values a un impact non significatif sur le résultat du modèle. Cela s'explique par la trop faible variation des valeurs de marché sur la base d'apprentissage. En effet, lors de la construction des arbres, les modèles séparent les scénarios avec des valeurs de marché qui dépassent un certain seuil. L'impact s'approche d'une classification et le montant exact des VM n'a pas d'importance. Cela confirme l'absence de prise en compte de l'effet taille.

## 4.3 XGboost

Enfin, le dernier modèle exploité est le modèle XGboost afin de prédire les mêmes indicateurs que dans les sections précédentes. Ces modèles de *machine learning* ont la réputation de donner des résultats très performants, qui peuvent surpasser ceux de *random forest*.

### Choix des paramètres

4 hyper-paramètres sont considérés pour la construction des modèles :

- nrounds qui correspond au nombre de fois où le *boosting* sera appliqué
- max.depth qui correspond à la profondeur maximale de l'arbre
- colsample\_bytree qui correspond au ratio de variables à utiliser dans la construction de chaque arbre
- eta qui correspond à la vitesse d'apprentissage.

La base de données  $T_0$  a été séparée en deux, une base d'apprentissage qui permet de calibrer les paramètres et de construire le modèle, et une base de test qui permet d'apprécier l'efficacité du modèle sur un échantillon de données qu'il ne connaît pas. La base d'apprentissage représente 80% des individus de notre base de données, après tirage aléatoire, et la base de test représente les 20% restants. Il est à noter que la calibration a été effectuée en suivant un algorithme de validation croisée à 5 blocs. Cet algorithme est présenté dans le chapitre de présentation des forêts aléatoires.

### nrounds

Sur la base d'apprentissage, la sélection de nrounds a été réalisée en traçant le RMSE de la forêt en fonction du nombre de *boosting* avec les paramètres par défaut. Le graphique est similaire pour les deux PVFP et le SCR et ressemble au graphique suivant :

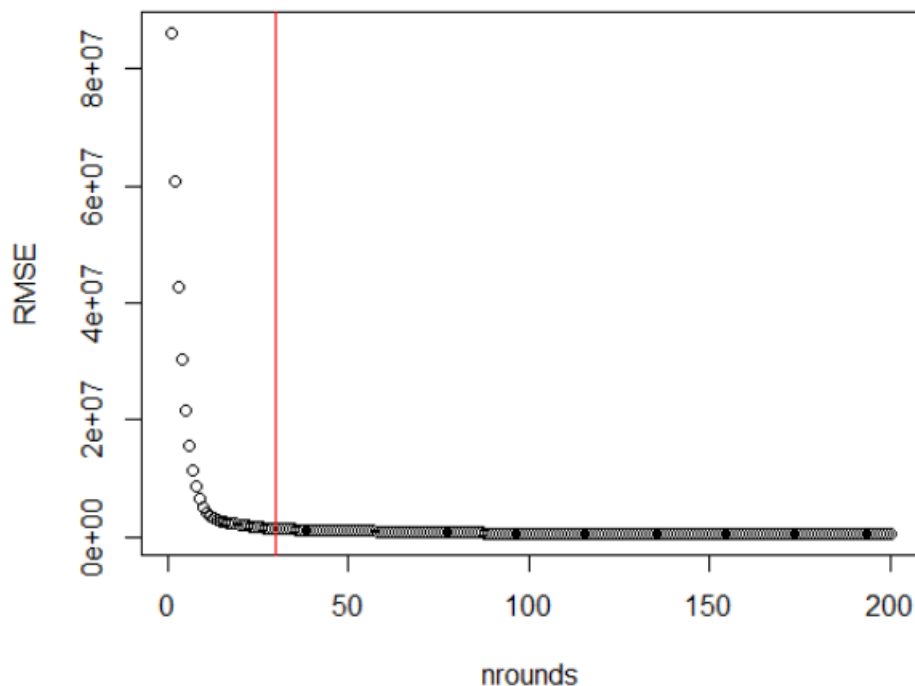


Figure 53 - XGboost : sélection du paramètre nrounds en fonction du RMSE

Cela permet de définir que l'erreur se stabilise à partir de 30 arbres en général et donc de fixer ce paramètre à 30.

### max.depth, colsample\_bytree et eta

Pour le choix de ces paramètres, une grille de recherche a été effectuée avec comme critère de sélection, la minimisation du RMSE. Les paramètres testés sont les suivants :

Paramètre	Valeur
max.depth	2   3   4   5
colsample_bytree	0,4   0,5   0,6   0,7   0,8
eta	0,1   0,2   0,3   0,4   0,5

Tableau 23 - Paramètres considérés pour XGboost

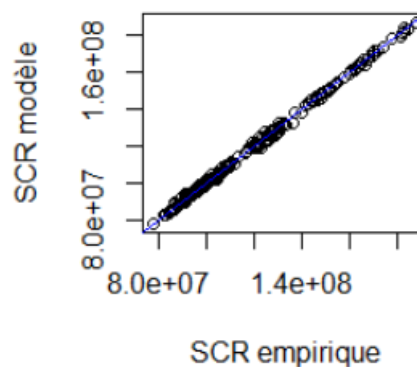
Les trop grandes valeurs pour ces paramètres risquent d'entraîner un sur-apprentissage, c'est pourquoi les valeurs testées maximales peuvent sembler éloigner du maximum possible. Les résultats pour chacun des modèles permettent de déterminer les paramètres retenus. Nous avons donc :

Modèles	Nrounds	max.depth	colsample_bytree	eta
SCR	30	5	0,7	0,4
PVFP RR	30	5	0,8	0,5
PVFP RN	30	5	0,8	0,3
PVFP RR/SCR	30	5	0,8	0,5
PVFP RN/SCR	30	5	0,8	0,3

Tableau 24 - Paramétrage des modèles XGboost

### Résultats $T_0$

Les modèles, créés sur la base d'apprentissage avec les paramètres précédemment déterminés, ont ensuite été appliqués sur la base de test. La représentation des valeurs théoriques et empiriques des différents indicateurs est la suivante :



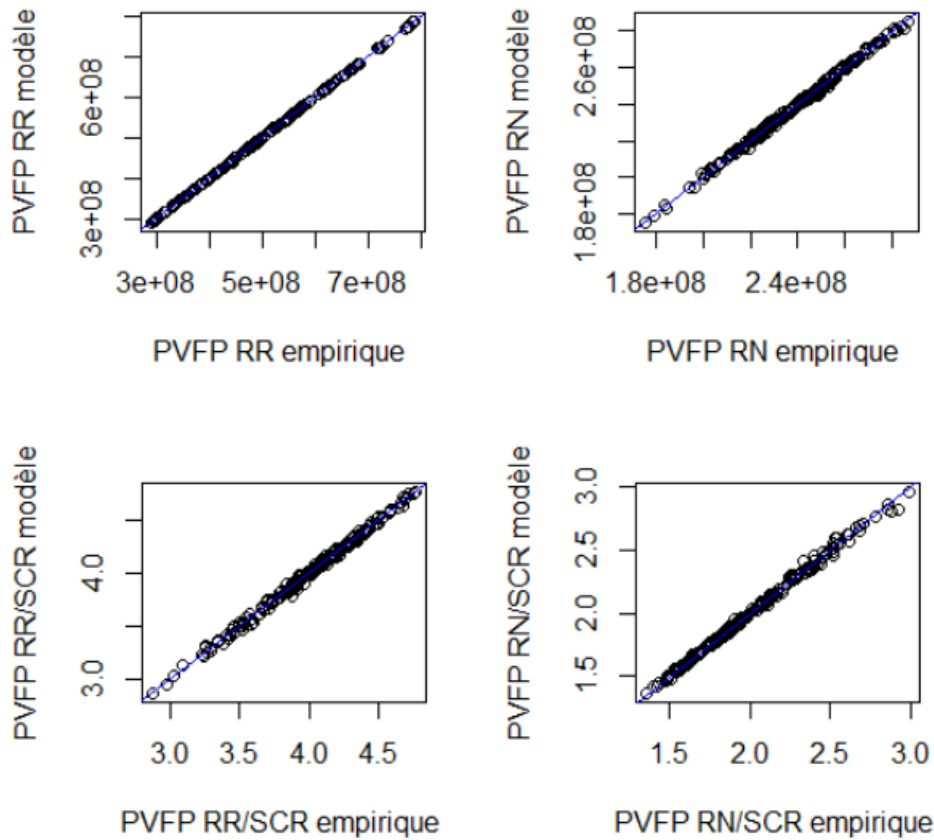


Figure 54 - XGboost : représentation des indicateurs théoriques et calculés

Les prédictions sont très proches des valeurs empiriques. L'erreur moyenne définie pour les modèles précédents est à nouveau utilisée, telle que donnée par l'expression :

$$\text{Erreur moyenne} = \frac{RMSE}{\bar{Y}}$$

Avec  $\bar{Y}$  la moyenne de l'indicateur de la base de test. Une étude des résultats sur 1 000 modèles permet d'obtenir les statistiques suivantes :

Modèles	Erreur moyenne	min	max
SCR	0,76 %	0,48 %	1,19 %
PVFP RR	0,19 %	0,12 %	0,45 %
PVFP RN	0,50 %	0,36 %	0,68 %
PVFP RR/SCR	0,66 %	0,45 %	1,06 %
PVFP RN/SCR	0,96 %	0,69 %	1,48 %

Tableau 25 - Performance de XGboost en  $T_0$



## Résultats $T_1$

Les modèles ont été testés sur la base issue du *Portefeuille  $T_1$*  sans intégration de nouveaux contrats. Les résultats obtenus sur 1 000 modèles sont les suivants :

Modèles	Erreur moyenne	min	max
SCR	3,46 %	2,46 %	6,81 %
PVFP RR	5,27 %	3,83 %	6,81 %
PVFP RN	6,37 %	3,71 %	7,93 %
PVFP RR/SCR	6,75 %	6,02 %	7,41 %
PVFP RN/SCR	10,01 %	8,22 %	11,37 %

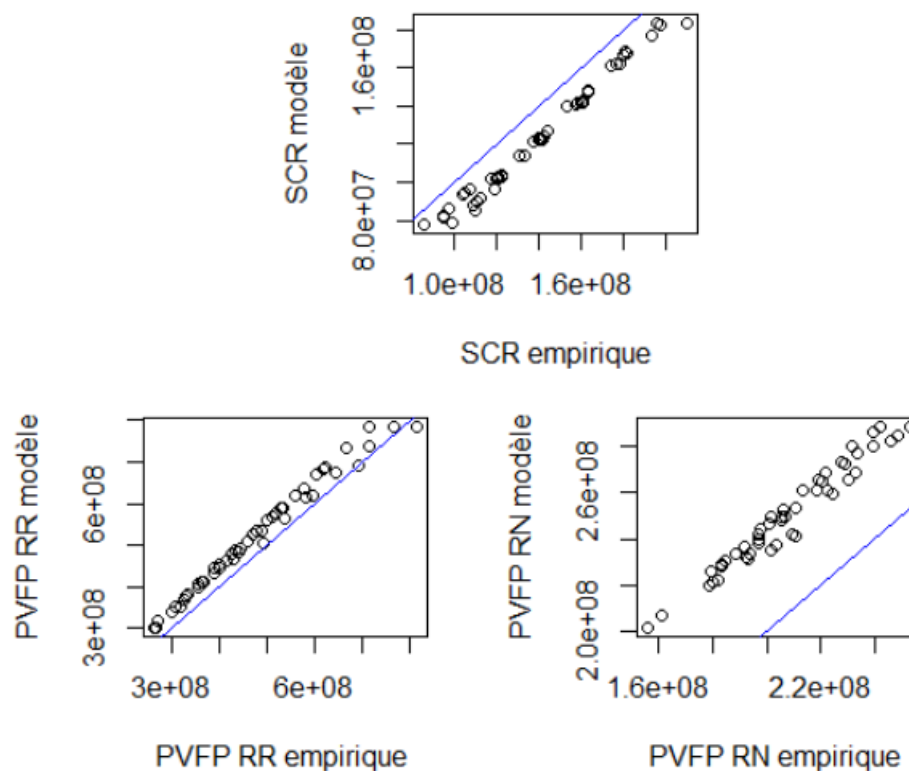
Tableau 26 - Performance de XGboost en  $T_1$

Les résultats sont beaucoup moins bons après le vieillissement, l'erreur moyenne des modèle SCR et PVFP RR sont semblables aux résultats de la *random forest*, néanmoins les autres indicateurs sont mal estimés. Le même exercice est réalisé avec la base de données issue du *Portefeuille  $T_1$*  avec intégration de nouveaux contrats. Les statistiques des modèles sur cette nouvelle base sont :

Modèles	Erreur moyenne	min	max
SCR	11,98265 %	9,52 %	15,37 %
PVFP RR	10,15535 %	8,92 %	11,89 %
PVFP RN	20,61178 %	18,32 %	22,13 %
PVFP RR/SCR	25,91111 %	22,38 %	28,12 %
PVFP RN/SCR	40,33834 %	34,73 %	43,36 %

Tableau 27 - Performance de XGboost en  $T_1$  après intégration de nouveaux contrats

Les modèles ne sont pas exploitables. La représentation des valeurs théoriques et empiriques des différents indicateurs permet d'observer un biais important. Les graphiques sont les suivants :



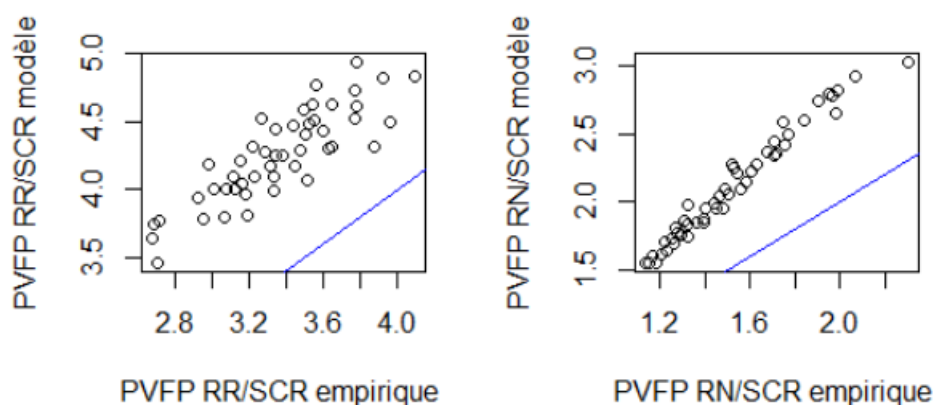


Figure 55 – Xgboost : représentation des indicateurs théoriques et calculés T1 avec nouveaux contrats

Le même type de biais que pour les *random forest* est observé. Comme pour ces derniers, ils s’expliquent en partie par une forte variation des VM à la suite du vieillissement intégrant de nouveaux contrats, qui n’est pas prise en compte par les modèles. Le test de sensibilité aux plus-values permettra de confirmer ou non cette hypothèse.

#### Test de sensibilité aux plus-values

Le test est réalisé sur le scénario central du *Portefeuille T<sub>1</sub>* sans nouveaux contrats pour le calcul de la PVFP RR. Les variables considérées sont les suivantes :

Taux cible action	9,5%
Taux cible immobilier	7,7%
Taux cible <i>cash</i>	6%
Taux cible obligations	76,8 %
Taux plus-values latente actions	21,5 %
Taux plus-values latente immobilier	14,4 %
Ratio PPE / PM	5,7 %
VM action	401 475 222
VM immobilier	326 899 113
VM <i>cash</i>	231 651 746
VM obligations	3 252 916 481

Tableau 28 - Caractéristiques du scénario initial après vieillissement du portefeuille

Comme pour les *random forest*, 1000 modèles ont été lancés sur ce scénario en situation de hausse ou de baisse des plus-values actions et immobilier, tel que détaillé dans le tableau ci-dessous. Pour chacune des situations suivantes, les statistiques obtenues sur la PVFP RR calculée sont :

Situation	Moyenne	min	max
-75 % PMVL action et immobilier	430 055 332	418 325 856	437 116 928
-50 % PMVL action et immobilier	430 055 332	418 325 856	437 116 928
-25 % PMVL action et immobilier	430 055 332	418 325 856	437 116 928
0 % PMVL action et immobilier	430 055 332	418 325 856	437 116 928
25 % PMVL action et immobilier	433 388 426	419 771 072	439 819 488
50 % PMVL action et immobilier	433 388 426	419 771 072	439 819 488
75 % PMVL action et immobilier	433 388 426	419 771 072	439 819 488
100 % PMVL action et immobilier	433 388 426	419 771 072	439 819 488
125 % PMVL action et immobilier	433 388 426	419 771 072	439 819 488
150 % PMVL action et immobilier	433 388 426	419 771 072	439 819 488
175 % PMVL action et immobilier	433 388 426	419 771 072	439 819 488
200 % PMVL action et immobilier	433 388 426	419 771 072	439 819 488

Tableau 29 - Test de sensibilité aux PMVL de XGboost

Encore une fois, les modèles ne prennent presque pas en compte la variation des plus-values. Cela s'explique par la trop faible variation des valeurs de marché sur la base d'apprentissage.

## 4.4 Comparaison des modèles

Les tableaux suivants récapitulent l'erreur moyenne de chaque modèle, sur chacun des portefeuilles :

Modèles	GLM $T_0$	Random forest $T_0$	XGboost $T_0$
SCR	6,17 %	2,97 %	0,76 %
PVFP RR	1,53 %	1,84 %	0,19 %
PVFP RN	2,61 %	2,26 %	0,50 %
PVFP RR/SCR	5,32 %	2,89 %	0,66 %
PVFP RN/SCR	5,48 %	4,21 %	0,96 %

Tableau 30 - Récapitulatif de la performance des modèles en  $T_0$

Modèles	GLM $T_1$	Random forest $T_1$	XGboost $T_1$
SCR	7,98 %	3,39 %	3,46 %
PVFP RR	5,10 %	4,77 %	5,27 %
PVFP RN	4,85 %	3,08 %	6,37 %
PVFP RR/SCR	10,78 %	4,65 %	6,75 %
PVFP RN/SCR	10,83 %	6,62 %	10,01 %

Tableau 31 - Récapitulatif de la performance des modèles en  $T_1$

Modèles	GLM $T_1$ contrats	Random forest $T_1$ contrats	XGboost $T_1$ contrats
SCR	26,23 %	10,83 %	11,98 %
PVFP RR	11,92 %	8,81 %	10,15 %
PVFP RN	26,82 %	16,68 %	20,61 %
PVFP RR/SCR	48,82 %	22,42 %	25,91 %
PVFP RN/SCR	69,89 %	33,65 %	40,33 %

Tableau 32 - Récapitulatif de la performance des modèles en  $T_1$  avec intégration de nouveaux contrats

Comme on peut l'observer, les Forêts aléatoires se démarquent pour leurs résultats en  $T_1$ . Il est noté tout de même que les *random forest*, ainsi que les modèles XGboost prennent difficilement en compte les valeurs de marché. En effet, comme les tests de sensibilité ont pu le démontrer, seule la GLM permet de prendre en compte cette variable avec la base de données qui a été utilisée. L'intégration de nouveaux contrats après vieillissement du portefeuille ne permet néanmoins pas, avec les hypothèses actuelles, d'estimer les différents indicateurs de façon pertinente.

De plus, la *random forest* a l'avantage d'être facile à calibrer par opposition au modèle XGboost. En effet, celle-ci nécessite la calibration de peu de paramètres afin d'être efficace et évite plus facilement le phénomène de sur-apprentissage. Dans le cas de XGboost, beaucoup de paramètres entrent en jeu et le sur-apprentissage arrive fréquemment.

Par ailleurs, la génération de la base de données par palier d'allocation, bien que cette méthode soit souvent utilisée par les assurances, est une approche qui semble moins efficace qu'une génération aléatoire pour répondre à notre problématique. La génération aléatoire aurait permis d'avoir une distribution plus cohérente dans le cadre de la GLM, et aurait pu permettre d'éviter que les forêts aient une approche qui s'apparente plus à une classification, qu'à une régression.

La démarche pourrait éventuellement être améliorée par l'utilisation de plus de variables, néanmoins cela entraînerait un temps de calcul exponentiellement plus long pour la génération de la base de données issue du modèle ALM. La constitution de la base utilisée a impliqué un temps de calcul de 14 jours pour les presque 2 000 calculs nécessaires, malgré l'utilisation de techniques de

parallélisation des calculs sous Excel. A titre de comparaison, l'utilisation des modèles construits pour 1 000 scénarios s'effectue en quelques secondes.

Finalement, l'utilisation d'un modèle de *random forest* permet de représenter les PVFP RR en fonction du SCR, après une année de vieillissement et sans intégration de nouveaux contrats, de la façon suivante :

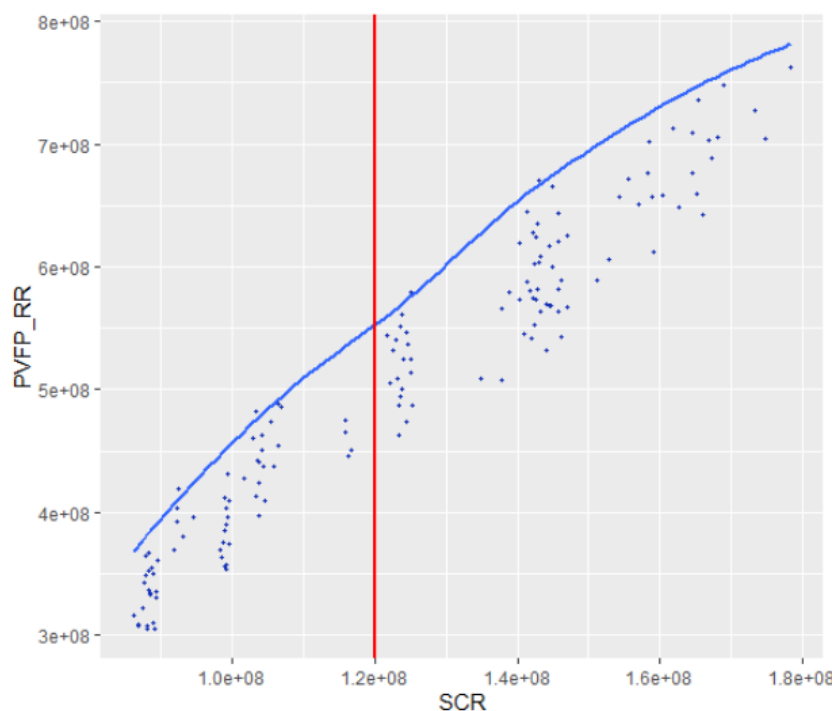


Figure 56 - Estimation par random forest de la PVFP RR et du SCR en T1 sans intégration de nouveaux contrats

Un assureur cherchant à déterminer l'allocation optimale de ses actifs pour un niveau de SCR donné pourra ainsi utiliser le modèle. A titre d'illustration, pour un niveau de SCR de 120 millions maximum qui est représenté par la droite verticale en rouge sur le repère, l'assureur qui représente notre portefeuille initial retiendra l'allocation liée aux résultats (en millions) :

Taux action	Taux immobilier	Taux cash	Taux obligations	SCR	PVFP RR
17 %	6 %	0 %	77 %	106,38	488,03

Tableau 33 - Allocation optimale pour un niveau de SCR de 120 millions

# Conclusion

Dans le cadre de ce mémoire, l'objectif des travaux présentés est d'évaluer la pertinence de l'utilisation du *machine learning* comme méthode alternative à celles existantes pour déterminer le choix de l'allocation optimale des actifs d'un assureur. Ces travaux s'inscrivent dans le contexte de Solvabilité II.

Pour ce faire, la démarche suivie pour la construction de la base de calcul est la suivante. Tout d'abord, un portefeuille d'épargne a été construit à partir de diverses hypothèses. Une base de données initiale en  $T_0$  a ensuite été générée en projetant ce dernier avec un modèle ALM, sur un ensemble de points d'allocation. Les métriques utilisées afin d'évaluer chacune des allocations testées sont le SCR de marché et la PVFP monde réel de la part des fonds en euros du portefeuille. Par la suite, deux vieillissements du portefeuille sur 1 an ont été réalisés afin d'évaluer les performances des modèles de *machine learning* après 1 an de vie du portefeuille. La différence entre ces deux vieillissements est l'intégration de nouveaux contrats dans l'un d'entre eux. Cela a permis finalement d'avoir à disposition 3 bases de données :

- La première, en  $T_0$ , qui sert à la construction et au calibrage des modèles de *machine learning*
- La seconde, en  $T_1$ , qui permet d'évaluer la performance des modèles après vieillissement du portefeuille
- La troisième, en  $T_1$  et avec intégration de nouveaux contrats, qui permet d'étudier l'efficacité des modèles dans un cadre plus proche de l'évolution attendue du portefeuille d'un assureur

Par la suite, des modèles GLM, *random forest* et XGboost ont été construits sur la base en  $T_0$ , et évalués sur les bases en  $T_1$ , dans le but de déterminer le SCR de marché et la PVFP monde réel pour différentes allocations testées. Les *random forest* ont permis de prédire avec la plus faible erreur, qui est basée sur le RMSE, nos indicateurs. En effet, ces modèles ont montré une erreur moyenne de 3% pour le SCR et de 1,8% pour la PVFP monde réel sur la base en  $T_0$ . Les résultats en  $T_1$  restent concluants, avec 3,4% d'erreur pour le SCR et 4,8% pour la PVFP monde réel.

Néanmoins plusieurs limites sont apparues face à la démarche employée. Bien que les résultats soient très précis lors de l'application des modèles sur une base de test en  $T_0$ , et le restent en  $T_1$ , il est observé que l'intégration de nouveaux contrats après 1 an de vie du portefeuille entraîne un biais important dans les prédictions. Dans la réalité, le portefeuille d'un assureur va intégrer de nouveaux contrats au fil du temps, les modèles montrent une difficulté à les prendre en compte. Cela peut être expliqué en partie par la faible prise en compte de certaines variables tels que les valeurs de marchés, qui sont fortement liées aux SCR et PVFP. N'ayant pas beaucoup varié dans la base de données initiale, leur impact n'a pas pu être complètement capté dans la construction des modèles. Des tests de sensibilité aux variations des valeurs de marché ont été mis en place afin de confirmer cette problématique.

De plus, la méthode de génération de notre base de calcul en faisant varier nos allocations par paliers a un effet certain sur les modèles réalisés. En effet, l'estimation des différents indicateurs correspond à un problème de régression, or dans ce cadre, les paramètres devraient être des variables quantitatives continues. La génération par paliers implique que les paramètres sont des variables quantitatives discrètes, car peu de valeurs sont possibles pour l'ensemble des scénarios. Les modèles *random forest* et XGboost considèrent donc qu'il s'agit d'un problème de classification au lieu de régression. Ainsi, un changement trop faible entre 2 hypothèses ne permettra pas d'obtenir un résultat

différent. Cela entraîne notamment la non prise en compte des valeurs de marché, pour ces modèles, lors de nos tests de sensibilité.

Il est à noter que pour déterminer une allocation optimale dans le cadre de Solvabilité II, le SCR de marché n'est peut-être pas le seul indicateur de risque pertinent à prendre en compte. En effet, il peut être intéressant de considérer le SCR de souscription car il est fortement dépendant des hypothèses sur les actifs, notamment car ces dernières jouent un rôle sur les rachats de l'assureur. Dans le cadre des travaux présentés, cet indicateur n'est pas pris en compte pour des raisons opérationnelles ; l'application des chocs de souscription lors de la génération de la base de données aurait entraîné un temps de calcul trop long.

Enfin, il est intéressant de mentionner l'importance de l'explicabilité du modèle dans un cadre assurantiel. La mise en relation des valeurs prises par certaines variables et leur conséquence sur la prédiction est un enjeu dans différents domaines de l'actuariat. Les forêts aléatoires et les modèles XGboost sont limités par leur explicabilité, et souvent qualifiés de modèle « boîte noire ». Pour autant, ces derniers peuvent tout de même être utilisés en complément d'un modèle ALM afin de déterminer une allocation optimale d'actifs. Les modèles de *machine learning* sont infiniment plus rapides une fois construits. Un assureur pourrait effectuer une pré-sélection d'allocations retenues par ce biais, qui pourrait ensuite être confirmée par l'utilisation d'un modèle ALM sur ce nombre de points d'allocation, beaucoup plus faible que dans les pratiques actuelles.

Finalement, l'utilisation du *machine learning* est une méthode qui présente un réel intérêt. Son utilisation seule ne permet pas de répondre directement à la problématique posée étant donné le contexte assurantiel, mais son utilisation en tant que complément permet un gain de temps conséquent par rapport aux méthodes plus classiques.

# Lexique

**ACPR** : Autorité de Contrôle Prudentiel et de Résolution  
**ALM** : *Asset and Liability management*  
**BE** : *Best Estimate*  
**CART** : *Classification And Regression Trees*  
**CDS** : *Credit Default Swap*  
**CML** : *Capital Market Line*  
**EEA** : *European Economic Area*  
**EIOPA** : *European Insurance and Occupational Pensions Authority*  
**GBM** : *Gradient Boosting Modele*  
**GLM** : *Generalized Linear Model*  
**GSE** : Générateur de Scénario Economique  
**MAE** : *Mean Absolute Error*  
**MAPE** : *Mean Absolute Percentage Error*  
**MEDAF** : Modèle d'Evaluation des Actifs Financiers  
**MSE** : *Mean Square Error*  
**NAV** : *Net Asset Value*  
**OCDE** : Organisation de Coopération et de Développement Economiques  
**OPCI** : Organisme de Placement Collectif Immobilier  
**PB** : Participation aux Bénéfices  
**PE** : Participation aux Excédents  
**PM** : Provision Mathématique  
**PMVL** : Plus ou Moins-Values Latentes  
**PPB** : Provision pour Participation aux Bénéfices  
**PPE** : Provision pour Participation aux Excédents  
**PVFP** : *Present Value of Future Profit*  
**RC** : Réserve de Capitalisation  
**RMSE** : Root Mean Square Error  
**RN** : Risque Neutre  
**RR** : Risque Réel  
**SCI** : Société Civile Immobilière  
**SCPI** : Société Civile de Placement Immobilier  
**SCR** : *Solvency Capital Requirement*  
**SML** : *Security Market Line*  
**TMG** : Taux Minimum Garanti  
**UC** : Unités de Comptes  
**VaR** : *Value at Risk*  
**VM** : Valeur de Marché  
**VNC** : Valeur Nette Comptable



# Liste des figures

1 - Représentation d'une frontière efficiente .....	6
2 - Représentation d'une frontière efficiente après intégration de l'actif sans risque.....	7
3 - Représentation de la CML avec 3 titres fictifs.....	8
4 - Piliers de la réforme Solvabilité II.....	13
5 - Passage de l'actif sous Solvabilité II .....	14
6 - Passage du passif sous Solvabilité II .....	14
7 - Pieuvre du SCR .....	15
8 - Représentation du SCR en scénario choqué .....	15
9 - Fonctions clés du pilier 2 de Solvabilité II .....	16
10 - Matrice de corrélation du SCR de marché .....	18
11 - Matrice de choc de spread pour le SCR bonds.....	21
12 - Matrice de choc de spread pour le SCR securisation .....	21
13 - Matrice de choc de spread pour le SCR cd.....	22
14 - Représentation d'un plan de solution sur 3 dimensions issu du site : <a href="https://jcrisch.wordpress.com/2015/04/02/les-reseaux-de-neurones/">https://jcrisch.wordpress.com/2015/04/02/les-reseaux-de-neurones/</a> .....	24
15 - Illustration de l'opérateur de sélection.....	26
16 - Illustration de l'opérateur de croisement .....	26
17 - Illustration de l'opérateur de mutation .....	26
18 - Démarche d'un algorithme génétique .....	26
19 - Démarche d'un projet de machine learning .....	28
20 - Classification d'individus lors d'un apprentissage non-supervisé.....	30
21 - Représentation du sur-apprentissage.....	31
22 - Représentation de la structure d'un arbre de décision .....	35
23 - Représentation du processus d'une forêt aléatoire .....	37
24 - Approche des algorithmes CART, random forest et XGboost par rapport aux observations .....	39
25 - Architecture du modèle ALM inspiré d'une note technique Optimind .....	44
26 - Projection de l'action et de l'immobilier en moyenne.....	51
27 - Bilan en norme comptable .....	56
28 - Situation initiale des actifs .....	56
29 - Caractéristiques du passif .....	57
30 - Bilan sous Solvabilité II .....	58
31 - Extrait de la base de données .....	60
32 - Représentation du SCR et de la PVFP RR dans la base.....	60
33 - Représentation du SCR et de la PVFP RN dans la base .....	61
34 - Densités de la base de données .....	62
35 - Matrice des corrélations .....	62
36 - Représentation des allocations pour la PVFP RR .....	63
37 - Représentation des allocations pour la PVFP RN.....	63
38 - Caractéristiques du passif du portefeuille $T_1$ .....	64
39 - Caractéristiques du passif du portefeuille $T_1$ avec nouveaux contrats .....	65
40 - GLM : Représentation de la densité et des fonctions de répartitions pour chaque indicateur ....	67
41 - GLM : représentation des indicateurs théoriques et calculés .....	69
42 - GLM : représentation des résidus de Pearson pour la PVFP RR.....	70
43 - GLM : sensibilité aux PMVL .....	71
44 - Random Forest : sélection du paramètre ntree en fonction du MSE .....	72
45 - Random Forest : sélection du paramètre mtry en fonction du RMSE pour le SCR.....	73

46 - Random Forest : sélection du paramètre maxnodes en fonction du RMSE pour le SCR.....	73
47 - Random Forest : sélection du paramètre mtry en fonction du RMSE pour la PVFP RR.....	74
48 - Random Forest : sélection du paramètre mtry en fonction du RMSE pour le ratio PVFP RR/SCR	74
49 - Random Forest : représentation des indicateurs théoriques et calculés .....	75
50 - Random Forest : ordre d'importance des variables pour le SCR .....	76
51 - Random Forest : ordre d'importance des variables pour la PVFP RN.....	76
52 - Random Forest : représentation des indicateurs théoriques et calculés T1 avec nouveaux contrats .....	78
53 - XGboost : sélection du paramètre nrounds en fonction du RMSE .....	80
54 - XGboost : représentation des indicateurs théoriques et calculés .....	82
55 - Xgboost : représentation des indicateurs théoriques et calculés T1 avec nouveaux contrats .....	84
56 - Estimation par random forest de la PVFP RR et du SCR en T1 sans intégration de nouveaux contrats .....	87

# Liste des tableaux

1 - Hypothèse des actifs .....	56
2 - Hypothèse des actifs, part d'obligations .....	57
3 - Hypothèses constantes des model points .....	58
4 - Caractéristiques des model points .....	58
5 - Décomposition du SCR de marché pour le scénario initial .....	59
6 - Indicateurs principaux pour le scénario initial .....	59
7 - Valeurs possibles pour chaque paramètre considéré .....	60
8 - Hypothèse des actifs du portefeuille $T_1$ .....	64
9 - Hypothèse des actifs du portefeuille $T_1$ , part des obligations .....	64
10 - Hypothèse des actifs du portefeuille $T_1$ avec nouveaux contrats .....	65
11 - Hypothèse des actifs du portefeuille $T_1$ avec nouveaux contrats, part des obligations .....	65
12 - Performance de la GLM en $T_0$ .....	69
13 - Performance de la GLM en $T_1$ .....	70
14 - Performance de la GLM en $T_1$ avec nouveaux contrats .....	70
15 - Caractéristiques du scénario initial après vieillissement du portefeuille .....	71
16 - Test de sensibilité aux PMVL de la GLM .....	71
17 - Paramétrage des modèles random forest .....	74
18 - Performance de la random forest en $T_0$ .....	76
19 - Performance de la random forest en $T_1$ .....	77
20 - Performance de la random forest en $T_1$ après intégration de nouveaux contrats .....	77
21 - Caractéristiques du scénario initial après vieillissement du portefeuille .....	78
22 - Test de sensibilité aux PMVL de la random forest .....	79
23 - Paramètres considérés pour XGboost .....	81
24 - Paramétrage des modèles XGboost .....	81
25 - Performance de XGboost en $T_0$ .....	82
26 - Performance de XGboost en $T_1$ .....	83
27 - Performance de XGboost en $T_1$ après intégration de nouveaux contrats .....	83
28 - Caractéristiques du scénario initial après vieillissement du portefeuille .....	84
29 - Test de sensibilité aux PMVL de XGboost .....	85
30 - Récapitulatif de la performance des modèles en $T_0$ .....	86
31 - Récapitulatif de la performance des modèles en $T_1$ .....	86
32 - Récapitulatif de la performance des modèles en $T_1$ avec intégration de nouveaux contrats .....	86
33 - Allocation optimale pour un niveau de SCR de 120 millions .....	87

# Bibliographie

Bonnefoy P. [2016] « Implémentation et calibrage d'un Générateur de Scénarios Economiques : impact sur la volatilité du *Solvency Capital Requirement* »

Choquer R. [2016] « Allocation stratégique d'actifs sous contrainte Solvabilité II »

De Lignaud M. [2018] « Comparaison de modèles prédictifs pour l'évaluation des coûts matériels automobiles »

Doullaye I. [2016] « Etude sous Solvabilité 2 d'un contrat d'assurance-vie en unités de compte avec garantie plancher en cas de décès et en cas de vie »

Franquet S. [2018] « Modélisation de la fréquence des sinistres graves en assurance automobile : apports et interprétabilité des méthodes d'apprentissage statistique »

Guillot A. [2015] « Apprentissage statistique en tarification non-vie : quel avantage opérationnel ? »

Groupe de travail *Best Estimate Liabilities Vie* [2016] « Exemples de pratiques actuarielles applicables au marché français »

Jacques L., Rain E. [2015] « Du modèle GLM à une approche darwinienne : Nouvelle génération de concepts et d'indicateurs pour l'optimisation du renouvellement Auto »

Jacquinet A. [2015] « Modélisation du Best Estimate et des provisions sous Solvabilité 2 du portefeuille épargne d'Etika »

Karamoko C. [2017] « Approche tarifaire des contrats collectifs Frais de Santé à l'aide des méthodes d'apprentissage »

Kouo K. [2017] « Tarification automobile : GLM vs réseaux de neurones »

Levy A. [2017] « Solvabilité II : Exigences quantitatives et impact comptables sur une société d'assurance mutuelle non-vie »

Ochoa Magana J. [2020] « Analyse de la reconnaissance de l'état de catastrophe naturelle à l'aide de la *Data science* et de l'*Open data* »

Ottou P. [2017] « Méthodes d'apprentissage automatique appliquées au provisionnement ligne à ligne en assurance non-vie »

Razafindrabary B. [2021] « ORSA et calcul prospectif du SCR par Machine Learning »

Tambrun H. [2020] « Allocation stratégique d'actifs en épargne dans le cadre d'une remontée rapide des taux d'intérêts »

Tinkey Ngatchou J. [2017] « Modélisation de la garantie incendie multirisque habitation, méthode paramétrique vs méthodes non-paramétriques »

Zouggagh F. [2018] « Tarification automobile à l'aide de modèles de *machine learning* et apport de données télématiques »