

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

Par : Lucas BLANCHETON

Titre : Allocation stratégique d'actifs : une approche par *reinforcement learning*
pour l'ALM

Confidentialité : NON (Durée : 1 an 2 ans)

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

Membres présents du jury de Signature
l'Institut des Actuaires

.....
.....
.....

Membres présents du jury de
l'ISFA

.....
.....
.....

Entreprise : Optimind

Nom : ROBERT

Signature :



*Directeurs de mémoire en entre-
prise :*

Noms : Guillaume SGAZAN

Stanislas YAO

Signature :



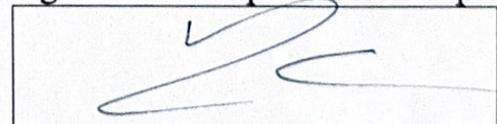
Invité :

Nom :

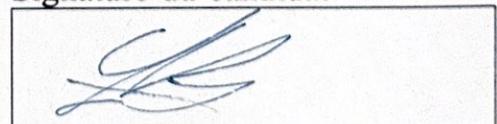
Signature :

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

Signature du responsable entreprise



Signature du candidat



Résumé

Dans un contexte économique au 31/12/2022 marqué par des taux d'intérêt élevés, les assureurs historiques font face à l'inertie de leur poche obligataire, ce qui les met en difficulté pour proposer des taux de valorisation de fonds euros alignés sur les taux de marché actuels. Cette étude vise à proposer une solution alternative à la stratégie d'allocation actuelle *Fixed-Mix* d'un assureur vie possédant à la fois des contrats euros et en unité de compte.

Ce travail explore l'application du *Reinforcement Learning* (RL), pour optimiser la stratégie d'allocation d'actifs dans le cadre de l'*Asset-Liability Management* (ALM) sous Solvabilité II fournissant un outil flexible dans le pilotage de sa stratégie d'allocation d'actifs en adéquation avec ses besoins. Dans un premier temps, après avoir introduit le cadre de travail dont l'ensemble des outils nécessaires à cette étude, l'algorithme de RL retenu, le *Deep Deterministic Policy Gradient* (DDPG), est implémenté dans le modèle ALM.

Quatre stratégies d'allocations d'actifs, élaborées à partir du DDPG, sont mises à l'épreuve, chacune visant à optimiser (maximiser ou minimiser selon leur nature) des métriques financières ou réglementaires telles que la *Present Value of Future Profits* (PVFP), le SCR de marché, le taux de rendement des actifs (TRA), et la richesse latente. Les stratégies sont évaluées en comparaison avec la stratégie *Fixed-Mix* de référence par rapport aux différentes métriques définies précédemment. L'étude inclut également une série d'analyses de sensibilité pour tester la robustesse et la flexibilité du modèle DDPG pour diverses configurations d'entraînement et de scénarios économiques. Ces analyses permettent d'examiner l'effet du nombre d'épisodes d'entraînement, des modifications apportées à la structure de récompense, et des sensibilités économiques sur la performance et la stabilité du modèle.

Il s'agira, à l'issue de l'ensemble de ces tests, de retenir l'allocation candidate la plus optimale, permettant une amélioration significative des métriques ciblées tout en démontrant une capacité d'adaptation aux fluctuations des conditions économiques et réglementaires.

Mots-clés : Reinforcement learning, Deep Deterministic Policy Gradient, ALM, stratégie d'allocation, Assurance-vie, Solvabilité II

Abstract

In an economic context as of 31/12/2022 marked by high interest rates, traditional insurers face the inertia of their bond portfolio, putting them at a challenge to offer Euro fund valuation rates aligned with the current market rates. This study aims to propose an alternative solution to the current Fixed-Mix allocation strategy of a life insurer owning both euro contracts and unit-linked contracts.

This work explores the application of *Reinforcement Learning* (RL) to optimize the asset allocation strategy within the Asset-Liability Management (ALM) framework under Solvency II, providing a flexible tool in steering the asset allocation strategy in line with its needs. Initially, after introducing the working framework and all the tools necessary for this study, the chosen RL algorithm, the Deep Deterministic Policy Gradient (DDPG), is implemented in the ALM model.

Four asset allocation strategies, provided from the DDPG, are tested, each aiming to optimize (maximize or minimize depending on their nature) financial or regulatory metrics such as the Present Value of Future Profits (PVFP), market risk, yield rate, and latent wealth. The strategies are evaluated in comparison with the reference Fixed-Mix strategy against the various metrics defined previously. The study also includes a series of sensitivity analyses to test the robustness and flexibility of the DDPG model for various training setups and economic scenarios. These analyses allow examining the effect of the number of training episodes, changes made to the reward structure, and economic sensitivities on the model's performance and stability.

At the end of all these tests, the most optimal candidate allocation will be retained, allowing for a significant improvement of the targeted metrics while demonstrating an adaptability to fluctuations in economic and regulatory conditions.

Keywords : *Reinforcement learning, Deep Deterministic Policy Gradient, ALM, asset allocation strategy, Life insurance, Solvency II*

Remerciements

J'exprime toute ma gratitude envers Christophe EBERLE, président fondateur d'Optimind part of Accenture, ainsi qu'aux partenaires de la practice Actuarial & Financial Services, Gildas ROBERT et Chloé PARFAIT, pour m'avoir offert l'opportunité d'effectuer mon stage et mon alternance au sein de leur entreprise. Un remerciement spécial également à Emmanuel BERTHELE pour sa lecture minutieuse et ses précieux conseils.

Mes remerciements les plus sincères vont à mes trois tuteurs, Guillaume SAGAZAN, Stanislas YAO et Marius MASSON, pour leur dévouement et leur bienveillance tout au long de la rédaction de ce mémoire. Leur disponibilité, leur approche pédagogique et leur soutien ont créé un environnement idéal qui a grandement contribué à l'aboutissement de ce travail.

Je tiens également à exprimer ma reconnaissance envers l'ensemble du corps professoral de l'ISFA pour m'avoir fourni le socle de connaissances nécessaire à l'aboutissement de ce mémoire. Un remerciement particulier à Denis CLOT pour son accompagnement, ses conseils avisés et ses relectures précieuses.

Je tiens également à remercier chaleureusement tous les consultants d'Optimind part of Accenture pour leurs conseils avisés, leur relecture et leur soutien tout au long de cette année. Plus particulièrement, toute la promotion d'alternants avec qui j'ai partagé des moments d'entraide et de partage : Geoffroy LAMBOLEZ, Killiann TANGUY, Annabel BERARD, Baptiste ALLAIRE, et Lisa RUELLAN.

Un grand merci à Amélie DE LA HAYE, Pablo GASSIOT, Aurélie AMARD, et Jorge OCHOA pour leur aide précieuse dans la relecture de ce mémoire.

Enfin, une pensée particulière pour ma famille pour leur soutien sans faille.

Table des matières

Résumé	i
Abstract	ii
Remerciements	iii
Synthèse	xi
Synthesis	xix
Introduction	xxvi
I Introduction au contexte réglementaire et modélisation de l'ALM	1
I.1 Spécificités de l'assurance vie	2
I.2 La réglementation Solvabilité II	3
I.3 Définitions et enjeux de l'ALM	9
I.4 Présentation du modèle ALM utilisé	10
I.4.1 Fonctionnement du modèle ALM	10
I.4.2 Description des mécanismes du modèle	11
I.5 Les générateurs de scénarios économiques	14
I.5.1 GSE risque neutre	14
I.5.2 GSE risque réel	16
I.6 Caractéristiques du portefeuille étudié	18
I.6.1 Le passif	18
I.6.2 L'actif	19
I.7 Définitions des indicateurs utilisés	21
I.8 Limites du modèle VBA dans le cadre actuel	24
I.9 Validation du modèle ALM	24
II La gestion actif-passif en assurance vie : état de l'art et lien avec le <i>Machine Learning</i>	27

II.1	Les différentes méthodes classiques de gestion	28
II.1.1	Les méthodes d'immunisation du portefeuille	30
II.1.2	Le modèle de Markowitz	31
II.1.3	Les modèles de surplus	34
II.1.4	Les modèles stochastiques	36
II.1.5	Les modèles dynamiques stochastiques	37
II.2	Le <i>reinforcement learning</i> pour la résolution d'un problème ALM	40
II.2.1	Concepts clés du <i>reinforcement learning</i>	40
II.2.2	Formalisation du processus de décision Markovien	42
II.2.3	Fonctions importantes	43
II.2.4	Équations de Bellman	44
II.2.5	Méthodes de résolution par renforcement : <i>Value-based</i>	44
II.2.6	Les réseaux de neurones	46
II.2.7	L'algorithme <i>Deep Q-Network</i>	51
II.2.8	Optimisation à l'aide du gradient de la politique	53
II.2.9	Le théorème du gradient de la politique	53
II.2.10	Les méthodes <i>Actor-Critic</i>	54
II.2.11	Gradient déterministe de la politique	55
III	Application de la méthode DDPG : étude et intégration dans le modèle ALM	57
III.1	L'algorithme <i>Deep Deterministic Policy Gradient</i>	58
III.2	Méthode de travail adoptée	62
III.3	Phase d'entraînement : approche et paramétrage	64
III.3.1	Configuration de l'espace d'états, d'actions et récompense	64
III.3.2	Calibrage du modèle	69
IV	Analyse et interprétation des résultats	71
IV.1	Répartition des actifs dans les stratégies d'allocation	72
IV.2	Impacts des résultats sur les différentes métriques en entraînement risque neutre	77
IV.2.1	Stratégie $SCR_{\text{marché}}$	77
IV.2.2	Stratégie PVFP	81
IV.2.3	Stratégie $SCR_{\text{marché}}$ et PVFP	85
IV.3	Impacts des résultats sur les différentes métriques en entraînement risque réel	88
IV.4	Analyse de sensibilité du modèle	90
IV.4.1	Sensibilité sur le nombre d'épisodes d'entraînement	90
IV.4.2	Sensibilité sur la structure de récompense du modèle	92
IV.4.3	Sensibilité sur la PVFP/ $SCR_{\text{souscription}}$	94
IV.5	Analyse de sensibilités économiques	98

IV.5.1 Impact des plus-values latentes obligataires sur le modèle	98
IV.5.2 Impact d'un changement de la courbe des taux de -100bps	103
Conclusion	112
Annexes	116
A.1 Matrices de corrélation $SCR_{marché}$	116
A.2 Définitions martingales et mouvement brownien	117
A.3 Portefeuille pour la frontière de Markowitz	117
A.4 L'algorithme <i>Q-learning</i>	118
A.5 Démonstration du Théorème du Gradient de la Politique	118

Table des figures

0.1	Interaction Agent/Environnement dans le cadre du <i>reinforcement learning</i>	xii
0.2	Description de la méthodologie	xiii
0.3	Pourcentages de la part des actions au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs	xiv
0.4	Pourcentages de la part des obligations au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs	xiv
0.5	Pourcentages de la part de l'immobilier au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs	xv
0.6	Agent/Environment Interaction in the context of <i>reinforcement learning</i>	xix
0.7	Description of the methodology	xxi
0.8	Percentage of equity in the portfolio according to different asset allocation strategies .	xxii
0.9	Percentage of bonds in the portfolio according to different asset allocation strategies .	xxii
0.10	Percentage of real estate in the portfolio according to different asset allocation strategies	xxiii
I.1	Courbes des taux EIOPA sans V.A	2
I.2	Piliers de la réforme Solvabilité II	4
I.3	Bilan Solvabilité II	5
I.4	Pieuvre du SCR	6
I.5	Calcul du SCR pour un risque donné	7
I.6	La dynamique ALM (Rodrigues Fontoura, [2020])	9
I.7	Fonctionnement du modèle ALM	10
I.8	Construction et fonctionnement d'un GSE	14
I.9	Courbe des taux sans risque fournie par l'EIOPA au 31/12/2022	15
I.10	Taux de PMVL initiales	20
I.11	Répartition initiale des classes d'actifs	20
I.12	Décomposition du SCR marché	25
I.13	Décomposition du SCR souscription	26
II.1	Frontière efficiente de Markowitz	33
II.2	Frontière efficiente et contrainte de déficit de Sharpe & Tint	35

II.3	Interaction Agent/Environnement dans le cadre du <i>reinforcement learning</i> selon Sutton et Barto [2017]	42
II.4	L'algorithme <i>Q-Learning</i>	45
II.5	Représentation d'un neurone du point de vue théorique selon (DISERBEAU, 2019)	46
II.6	Représentation d'une couche de neurones	47
II.7	Représentation d'un réseau de neurones	48
II.8	La <i>Backpropagation</i>	49
II.9	L'algorithme de descente de gradient	51
II.10	L'algorithme du <i>Deep Q-Learning</i>	52
II.11	L'algorithme <i>Actor-Critic</i>	54
III.1	L'algorithme <i>Deep Deterministic Policy Gradient</i>	58
III.2	Description de la méthodologie	63
III.3	Processus d'Ornstein-Uhlenbeck sur 40 ans	66
III.4	Description de la méthode d'entraînement du modèle	69
III.5	Exemple de l'évolution de la récompense moyenne pendant la calibration du modèle	69
IV.1	Pourcentages et statistiques de la part des actions au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs	73
IV.2	Pourcentages et statistiques de la part de l'immobilier au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs	74
IV.3	Pourcentages et statistiques de la part des obligations au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs	75
IV.4	Comparaison des métriques avec écart en % par rapport au scénario de référence	77
IV.5	Décomposition du SCR de marché et écart en % pour la stratégie d'optimisation $SCR_{marché}$ par rapport à la stratégie de référence	78
IV.6	Produits financiers en scénario central et choc de spread pour les deux stratégies	79
IV.7	Part en valeur de marché des obligations pour les deux stratégies	79
IV.8	TME et Taux servi réel pour la stratégie d'optimisation et de référence en scénario choc de taux up	79
IV.9	Montant de PPB pour la stratégie d'optimisation et de référence	80
IV.10	TME, Taux servi et réel pour la stratégie d'optimisation et de référence en scénario central	80
IV.11	VIF, RM ainsi que l'indicateur de solvabilité de l'allocation de référence et celle optimisant le SCR de marché	81
IV.12	Comparaison des métriques avec écart en % par rapport au scénario de référence	82
IV.13	Log-rendement des actions et de l'immobilier des projections risque neutre	83
IV.14	Décomposition du SCR de marché et écart en % pour la stratégie d'optimisation PVFP par rapport à la stratégie de référence	83

IV.15 VIF, RM ainsi que l'indicateur de solvabilité de l'allocation de référence et celle optimisant la PVFP	84
IV.16 Comparaison des métriques avec écart en % par rapport au scénario de référence . . .	85
IV.17 Décomposition du SCR de marché et écart en % pour la stratégie d'optimisation PVFP/SCR _{marché} par rapport à la stratégie de référence	86
IV.18 VIF, RM ainsi que l'indicateur de solvabilité de l'allocation de référence et celle optimisant le SCR de marché et la PVFP	87
IV.19 TRA pour la stratégie d'optimisation	88
IV.20 Richesse latente pour la stratégie d'optimisation	88
IV.21 Comparaison des métriques avec écart en % par rapport au scénario de référence . . .	89
IV.22 Temps de calcul en fonction du nombre d'épisodes d'entraînement	90
IV.23 PVFP en fonction du nombre d'épisodes d'entraînement	91
IV.24 SCR et SCR de marché en fonction du nombre d'épisodes d'entraînement	91
IV.25 Pourcentages des obligations au sein du portefeuille en fonction de la stratégie référence et PVFP/SCR _{souscription}	94
IV.26 Pourcentages des actions au sein du portefeuille en fonction de la stratégie référence et PVFP/SCR _{souscription}	94
IV.27 Pourcentages de l'immobilier au sein du portefeuille en fonction de la stratégie référence et PVFP/SCR _{souscription}	94
IV.28 Le SCR _{rachat} pour les deux stratégies	96
IV.29 Produits financiers dans le scénario de rachat massif pour les deux stratégies d'allocations	97
IV.30 TME, Taux servi des deux stratégies d'optimisation lors du scénario de rachat massif	97
IV.31 Taux de PMVL initiales dans une situation initiale de PVL obligataires	98
IV.32 Décomposition du SCR de marché pour la stratégie de référence	99
IV.33 Pourcentages des obligations au sein du portefeuille en fonction de la stratégie référence et PVFP/SCR _{marché}	99
IV.34 Pourcentages des actions au sein du portefeuille en fonction de la stratégie référence et PVFP/SCR _{marché}	99
IV.35 Pourcentages de l'immobilier au sein du portefeuille en fonction de la stratégie de référence et PVFP/SCR _{marché}	100
IV.36 Produits financiers générés pour les actifs de la stratégie référence	100
IV.37 Produits financiers générés pour les actifs de la stratégie PVFP/SCR _{marché}	100
IV.38 Comparaison des métriques avec écart relatif par rapport au scénario de référence . .	101
IV.39 Décomposition du SCR de marché et écart en % de la stratégie PVFP/SCR _{marché} par rapport à la stratégie de référence	102
IV.40 Impact du choc hausse des taux sur les produits financiers obligataires des deux stratégies	102
IV.41 Indicateur de solvabilité, RM ainsi que la VIF pour les deux stratégies	103
IV.42 Courbes des taux initiale et choquée de -100bps	104

IV.43 Décomposition du SCR marché dans le scénario de référence	104
IV.44 Pourcentages des obligations au sein du portefeuille pour les trois stratégies	105
IV.45 Pourcentages des actions au sein du portefeuille pour les trois stratégies	105
IV.46 Pourcentages de l'immobilier au sein du portefeuille pour les trois stratégies	105
IV.47 Comparaison des métriques avec écart par rapport au scénario de référence	106
IV.48 Produits financiers générés pour les actifs de la stratégie référence	106
IV.49 Produits financiers générés pour les actifs de la stratégie PVFP/SCR _{marché}	106
IV.50 Décomposition du SCR de marché et écart en % de la stratégie PVFP/SCR _{marché} par rapport à la stratégie de référence	107
IV.51 Écart de produits financiers actions entre le scénario central et choc action pour les stratégies	108
IV.52 Montant de loyer entre les deux stratégies	108
IV.53 Taux servi pour les deux stratégies en scénario central	109
IV.54 Taux servi pour les deux stratégies en scénario choc de taux à la baisse	109
IV.55 Taux servi pour les deux stratégies en scénario choc de taux à la hausse	110
IV.56 Indicateur de solvabilité, RM ainsi que la VIF pour les deux stratégies	110

Synthèse

Face à l'augmentation marquée des taux d'intérêt, avec le rendement moyen des fonds euros passant de 1,28 % en 2021 à 1,91 % en 2022 (selon les estimations de l'ACPR pour 2023, le taux est de 2,6 %) et le taux du Livret A grim pant de 0,5 % à 3 % entre 2022 et 2023, le secteur de l'assurance vie est poussé vers une révision stratégique en matière d'allocation d'actifs. L'étude se positionne au 31/12/2022, dans un environnement de taux d'intérêt élevés, où le portefeuille d'actifs affiche des plus-values latentes (PVL) pour ses composantes actions et immobilier, à l'exception de la poche obligataire. L'assureur concerné par cette étude propose des contrats multisupports euros et en unités de compte. Ce mémoire propose une solution alternative à la stratégie actuelle du modèle utilisé *Fixed-Mix* par une stratégie d'allocation dynamique basée sur le *reinforcement learning* (RL). Les méthodes de RL se distinguent par leur capacité à générer des propositions d'allocations d'actifs adaptatives, en prenant en compte les besoins spécifiques de l'assureur et les évolutions du marché. Cette allocation est construite dans un contexte économique de taux élevés où l'assureur, soumis à Solvabilité II, veut conserver un avantage concurrentiel et assurer une gestion prudente et optimale de ses engagements. Face à ce contexte, les assureurs traditionnels se heurtent au renouvellement lent de leur portefeuille obligataire, accumulé durant les périodes de taux bas. Cette contrainte limite leur aptitude à servir des taux de valorisation des fonds en euros qui reflètent les conditions actuelles du marché. Dans cette optique, le *reinforcement learning* apparaît comme une méthode adaptée à ce contexte, permettant d'offrir une gestion agile du portefeuille qui s'ajuste en temps réel aux évolutions du marché.

Le RL est une famille d'algorithmes d'apprentissage automatique où un agent apprend à prendre des décisions en interagissant avec un environnement, afin de maximiser une certaine métrique par le biais de la récompense comme l'illustre la figure 0.1 ci-dessous :

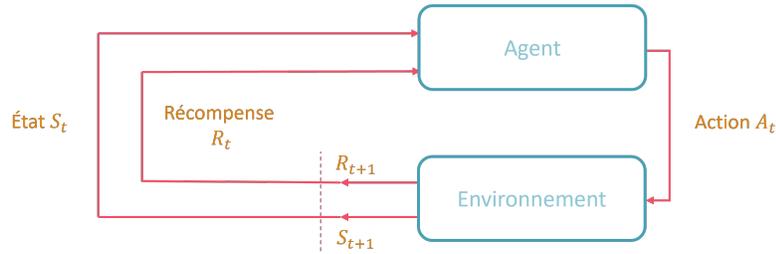


FIGURE 0.1 – Interaction Agent/Environnement dans le cadre du *reinforcement learning*

Dans le contexte de l'*asset liabilities management* (ALM), l'agent RL peut être considéré comme le gestionnaire de portefeuille qui décide des allocations de chacun des actifs.

Parmi les méthodes de *reinforcement learning*, le *Deep Deterministic Policy Gradient* (DDPG) est l'algorithme retenu et implémenté dans le modèle ALM au sein de cette étude. Le choix de ce modèle est lié au fait que l'algorithme est adapté aux problèmes d'optimisation dans les espaces d'actions continus. La conception unique du DDPG combine un *Actor* qui génère des actions à partir de l'état actuel de l'environnement, et le *Critic* qui évalue ces actions en calculant la valeur attendue des récompenses futures. Cette dualité favorise à la fois une exploration et une exploitation efficaces de l'environnement.

L'utilisation de l'algorithme DDPG s'articule autour de deux cadres distincts : un environnement risque neutre et un environnement risque réel, chacun défini par des objectifs propres d'optimisation qui s'appuient sur des métriques adaptées à leur contexte spécifique. Dans l'environnement risque neutre, l'accent est mis sur le $SCR_{\text{marché}}$ et la *present value of future profits* (PVFP), avec pour but de respecter les normes réglementaires tout en améliorant la performance financière de l'entreprise d'assurance. Dans l'environnement risque réel, les objectifs se concentrent sur le taux de rendement de l'actif (TRA) et la richesse latente. Ce couple permet d'obtenir une synergie intéressante entre le passif et l'actif pour l'assureur.

Une fois les métriques choisies dans chaque modèle, l'étape suivante consiste à mettre en oeuvre les modèles de DDPG dans chacun des univers de risque. L'objectif du DDPG est d'optimiser ces indicateurs, soit en les maximisant soit en les minimisant selon leur nature. Ainsi l'étude propose d'étudier quatre modèles chacun défini par des stratégies d'allocations d'actifs spécifiques :

- La stratégie PVFP (Maximiser la PVFP)
- La stratégie $SCR_{\text{marché}}$ (Minimiser le $SCR_{\text{marché}}$)
- La stratégie PVFP/ $SCR_{\text{marché}}$ (Maximiser la PVFP et minimiser le $SCR_{\text{marché}}$)
- La stratégie TRA/Richesse latente (Maximiser le TRA et maximiser la richesse latente)

Chaque modèle DDPG, après sa phase d'entraînement, détermine une allocation d'actifs optimale

selon les métriques définies. Ces stratégies optimisées grâce au modèle DDPG sont alors comparées à la stratégie existante, fondée sur une approche *Fixed-mix*. La méthodologie d'optimisation des stratégies adoptée est résumée dans le schéma 0.2 présenté ci-dessous :

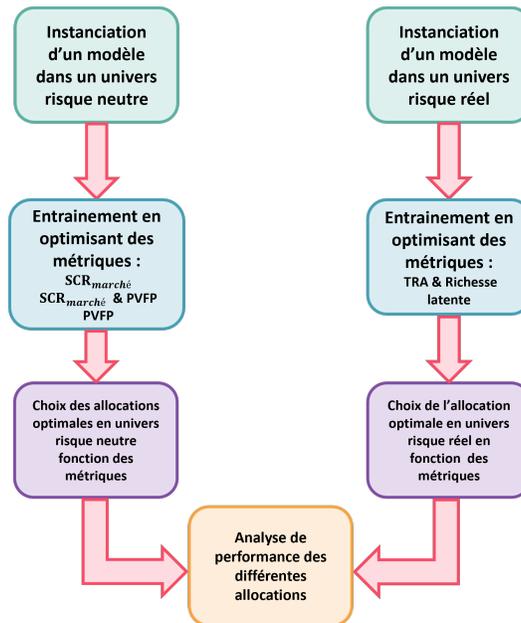


FIGURE 0.2 – Description de la méthodologie

La phase d'entraînement du DDPG dans le modèle ALM est une étape clé pour développer une stratégie d'allocation d'actifs optimisée. L'état de l'agent prend en compte la valeur des métriques en fin de projection en fonction de la stratégie étudiée (par exemple la PVFP et le $SCR_{marché}$ pour la stratégie PVFP/ $SCR_{marché}$), pour chaque indice (actions, immobilier et obligations) il comprend 40 valeurs correspondant aux données extraites du GSE et les 40 années de projection d'allocations d'actifs.

L'espace d'actions représente les pourcentages d'allocation possibles entre les différentes classes d'actifs. Des contraintes d'allocations sont également appliquées pour garantir la faisabilité et la conformité des stratégies, limitant par exemple les proportions d'investissement dans chaque classe d'actifs ainsi que le pourcentage de variation d'un actif d'un pas de temps à l'autre afin d'éviter les comportements brusques.

La fonction de récompense comme introduite par le schéma 0.1 est conçue pour orienter l'agent vers des actions optimales, en veillant à ce qu'il respecte l'ensemble des contraintes imposées. Au cours de la phase de calibration, le modèle a été soumis à un entraînement de 200 épisodes, chacun constitué de 1 000 simulations individuelles. Cette démarche méthodologique aboutit à un jeu de 200 000 trajectoires d'apprentissage. De plus, l'intégration d'un mécanisme de bruit, spécifiquement via le

processus d'Ornstein-Uhlenbeck, offre un moyen systématique d'encourager l'exploration évitant ainsi que l'agent ne stagne dans des optima locaux. Cette approche permet ainsi d'améliorer la robustesse du modèle en augmentant sa capacité de généralisation par le test d'allocations variées. À la fin des 200 000 trajectoires, le modèle observe quelle projection maximise sa récompense. La trajectoire sélectionnée définit alors la stratégie d'allocation d'actifs à adopter. Les allocations optimales par actifs, issues des simulations, sont détaillées ci-dessous :

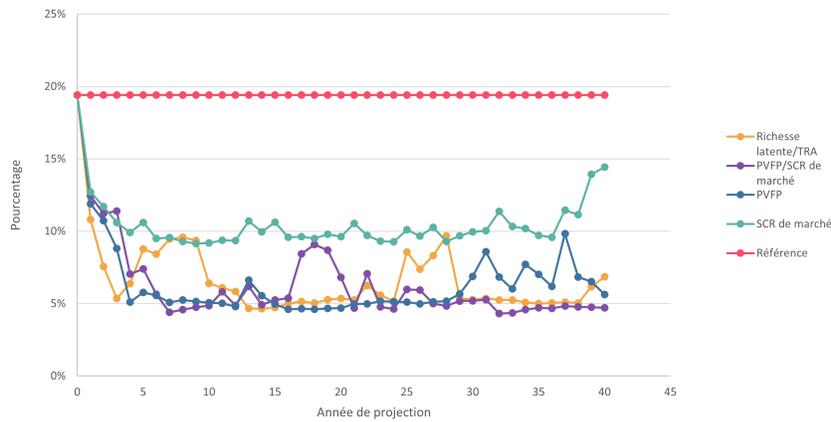


FIGURE 0.3 – Pourcentages de la part des actions au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs

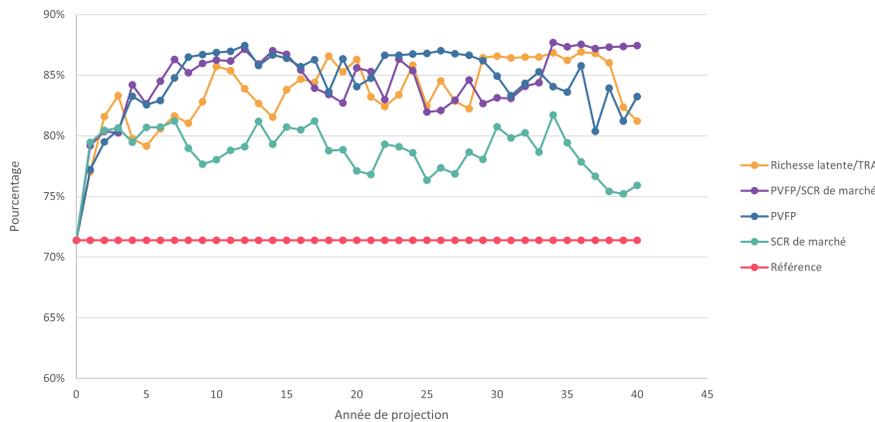


FIGURE 0.4 – Pourcentages de la part des obligations au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs

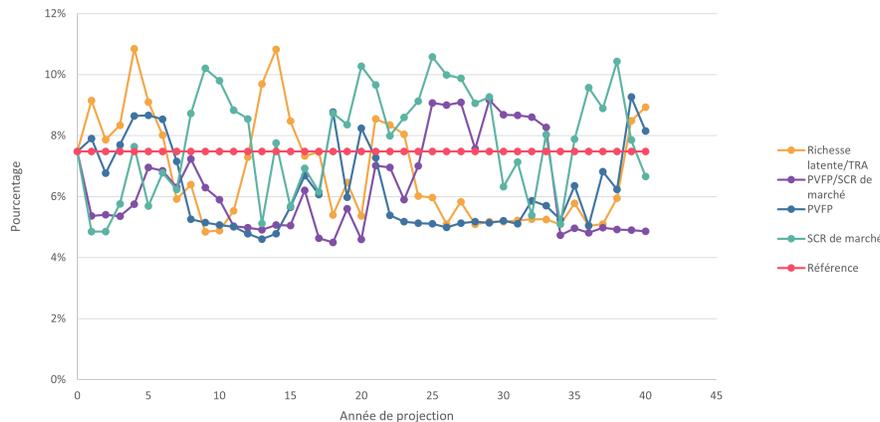


FIGURE 0.5 – Pourcentages de la part de l’immobilier au sein du portefeuille en fonction des différentes stratégies d’allocation d’actifs

Les figures ci-dessus (0.3, 0.4 et 0.5) illustrent la répartition par classes d’actifs des stratégies d’allocations optimales générées par le DDPG. Pour chacune des stratégies, le modèle évalue les métriques (PVFP, le $SCR_{\text{marché}}$ et le SCR) dans l’environnement risque neutre et les résultats sont présentés dans le tableau 0.1 ci-dessous. Cette démarche vise à fournir un point de comparaison avec les différentes stratégies fournies par le DDPG et la stratégie de référence *Fixed-Mix*.

L’algorithme révèle son efficacité en réalisant une amélioration notable de la PVFP. Spécifiquement, la stratégie axée sur la maximisation de la PVFP aboutit à une croissance de 13 % de cette métrique. De manière similaire, la stratégie combinant les objectifs de PVFP et d’une réduction du SCR de marché conduit à une hausse de 12 % de la PVFP, tout en permettant de diminuer le SCR de marché de 2 %. Enfin, l’approche centrée sur la minimisation du SCR de marché se traduit par une réduction de 4 % de cette même métrique. L’amélioration de la PVFP par les différentes stratégies d’allocations d’actifs est principalement attribuable aux *management actions* de l’algorithme consistant à réaliser massivement des PVL sur les actions en début de projection. Cette démarche est complétée par un réinvestissement dans de nouvelles obligations bénéficiant de taux d’intérêt plus élevés, contribuant ainsi directement à l’augmentation de la PVFP.

Stratégies	Référence	RN PVFP/ $SCR_{\text{marché}}$	RN PVFP	RN $SCR_{\text{marché}}$
Métriques				
PVFP	3 449 110	3 857 009 ↑	3 885 879 ↑	3 765 291 ↑
SCR	2 672 117	2 614 645 ↓	2 660 763 ↓	2 569 768 ↓
SCR Marché	1 955 249	1 938 986 ↓	1 988 921 ↑	1 874 316 ↓

TABLE 0.1 – Résumé de l’efficacité des stratégies sur les différentes métriques par rapport au scénario de référence

Après la première évaluation des stratégies avec un entraînement en univers risque neutre, la stratégie d'optimisation en univers d'entraînement risque réel consiste en l'amélioration du TRA et de la richesse latente. Les résultats obtenus montrent que l'algorithme a réussi à accroître significativement le TRA en vendant massivement les PVL actions et en achetant de nouvelles obligations, générant ainsi une PVFP de plus de 11 %. Cependant, cette stratégie s'est accompagnée d'une diminution de la richesse latente et d'une dégradation du SCR, rendant cette approche moins attractive dans un contexte réglementaire. Pour vérifier la capacité d'adaptation du modèle, une analyse de sensibilité est conduite.

L'analyse de sensibilité a pour but d'observer le comportement du modèle en ajustant divers paramètres, afin d'étudier leur influence sur la performance et la stabilité du système. Elle aborde deux aspects critiques pour évaluer la robustesse et l'adaptabilité du modèle DDPG : les sensibilités associées à sa calibration et les sensibilités associées aux conditions économiques. La stratégie optimisant le couple PVFP/SCR_{marché} est retenue pour tester ces impacts.

Pour les sensibilités de modèle, trois axes sont étudiés : la structure de récompense, le nombre d'épisodes d'entraînement et l'optimisation du couple PVFP/SCR_{souscription}. Pour la sensibilité du nombre d'épisodes d'entraînement le modèle est entraîné sur un éventail varié de nombre d'épisodes, allant de 50 à 250. L'objectif est de détecter le point où une augmentation du nombre d'épisodes ne se traduit plus par une amélioration significative des métriques, en considérant également les coûts computationnels associés. Les résultats de ce mémoire ont montré que le modèle de 200 épisodes apparaît comme le choix le plus optimal pour le couple optimisation métriques/temps de calcul.

La deuxième sensibilité liée au modèle étudié porte sur la récompense. Elle consiste à modifier la structure de calcul de celle-ci en introduisant des poids. Cette approche permet de réaliser un pilotage de l'optimisation d'une métrique par rapport à l'autre variable de l'étude. Pour cela, plusieurs combinaisons de poids ont été testées :

- "Référence" avec 0,5 de poids pour chaque métrique
- "Privilège PVFP" avec 0,75 pour la PVFP et 0,25 pour le SCR_{marché}
- "Privilège SCR_{marché}" avec 0,25 pour la PVFP et 0,75 pour le SCR_{marché}

Métrique Stratégie	SCR	SCR de marché	PVFP
Référence	2 614 645	1 938 986	3 857 009
Privilège PVFP	2 612 564 ↓	1 932 831 ↓	3 853 621 ↓
Privilège SCR de marché	2 585 805 ↓	1 903 632 ↓	3 855 377 ↓

TABLE 0.2 – Résultats de la sensibilité sur la structure de récompense

Le tableau 0.2 ci-dessus sur la sensibilité des poids affectés dans la fonction de récompense montre une

influence sur l'optimisation du SCR de marché, avec une réduction observée dans la stratégie privilégiant ce critère. Cependant, la stratégie focalisée sur la PVFP ne révèle qu'une légère baisse de cette dernière, suggérant un plateau dans son optimisation. Ceci indique que, la modification des poids peut effectivement améliorer le SCR de marché, tandis que l'impact sur la PVFP reste limité. Cela indique possiblement que l'optimisation de la PVFP a atteint un plateau dans ce contexte économique. Ainsi, malgré le fait de privilégier l'optimisation de la PVFP à travers l'ajustement des poids, le modèle n'a pas réussi à réaliser des améliorations significatives de cette métrique.

La dernière analyse de sensibilité du modèle est axée sur l'optimisation du couple PVFP/SCR_{souscription}. Cette sensibilité a pour but d'étudier l'impact d'une métrique plus étroitement liée au passif. Cependant, elle révèle une efficacité plus modeste. Si cette approche parvient à améliorer le SCR de souscription de -1 %, l'augmentation de la PVFP est relativement limitée, ne s'élevant qu'à +1 %. Ce résultat suggère que le modèle capte avec moins d'efficacité les interactions liées au passif.

Pour approfondir l'analyse, l'étude des sensibilités économiques se focalise sur deux aspects : l'impact des plus-values latentes obligataires en répliquant la situation économique au 31/12/2021 et les effets de l'impact de la courbe des taux d'un choc de -100 bps.

Métrique \ Sensibilité	SCR	SCR Marché	PVFP
PVL Obligataires	0 %	-3,6 %	+7,9 %
Courbe des Taux	-3,5 %	-5,1 %	+0,2 %

TABLE 0.3 – Résultats des sensibilités économiques par rapport à la stratégie *Fixed-Mix*

L'introduction de PVL obligataires couplée à un environnement économique au 31/12/2022 correspond à un contexte favorable pour l'assureur. Le tableau 0.3 illustre l'aptitude du modèle à exploiter les conditions de marché favorables pour améliorer la performance globale. Les résultats indiquent que le modèle capte efficacement les opportunités de PVL, témoignant de sa flexibilité et de son adaptation aux dynamiques de marché positives.

Toujours dans le tableau 0.3, la sensibilité relative au changement de la courbe des taux met en évidence un effet inverse sur la métrique privilégiée par le modèle. Confronté à une diminution soudaine des taux d'intérêt de -100bps, le modèle privilégie le SCR de marché par rapport à la PVFP. Ce résultat est en phase avec le nouveau contexte économique de la courbe des taux choquée de -100bps. En effet, la situation des taux d'intérêt plus faibles complique la libération de résultats financiers.

Ce travail a permis d'explorer l'efficacité du *reinforcement learning*, et plus spécifiquement du DDPG, pour proposer une stratégie d'allocation d'actifs. Les résultats encourageants de cette étude incitent

à continuer dans l'exploration du champ des possibles autour du *reinforcement learning*. Il est important de noter que, malgré ses avantages, l'algorithme peut entraîner des temps de calcul significatifs, nécessitant une attention particulière lors de son déploiement dans des environnements de production. Cet algorithme est construit pour fournir à l'assureur une méthode polyvalente et agile en fonction de ses besoins en matière de rendement ou de pilotage réglementaire. Le DDPG se présente comme un outil stratégique pour naviguer dans l'écosystème complexe de l'assurance, répondant aux défis posés par les fluctuations du marché et les cadres réglementaires évolutifs.

Synthesis

Given the marked increase in interest rates, with the average yield of Euro funds moving from 1.28 % in 2021 to 1.91 % in 2022 (according to ACPR estimates for 2023, the rate is 2.6%) and the rate of the Livret A climbing from 0.5 % to 3 % between 2022 and 2023, the life insurance sector is pushed towards a strategic revision in terms of asset allocation. The study is positioned as of 31/12/2022, in a high-interest rate environment, where the asset portfolio shows latent wealth for its equity and real estate components, except for the bond portfolio. The insurer concerned by this study offers multi-support euro and unit-linked contracts. This thesis proposes an alternative solution to the current model's Fixed-Mix strategy through a dynamic allocation strategy based on reinforcement learning (RL). RL methods stand out for their ability to generate adaptive asset allocation proposals, taking into account the specific needs of the insurer and market developments. This allocation is constructed in a high-interest rate economic context where the insurer, subject to Solvency II, wants to maintain a competitive advantage and ensure prudent and optimal management of its commitments. Given this context, traditional insurers face slow renewal of their bond portfolio, accumulated during periods of low rates. This constraint limits their ability to serve Euro fund valuation rates that reflect current market conditions. In this perspective, reinforcement learning appears as a method suited to this context, offering agile portfolio management that adjusts in real-time to market developments.

RL is a family of machine learning algorithms where an agent learns to make decisions by interacting with an environment, to maximize a certain metric through rewards, as illustrated in Figure 0.1 below :

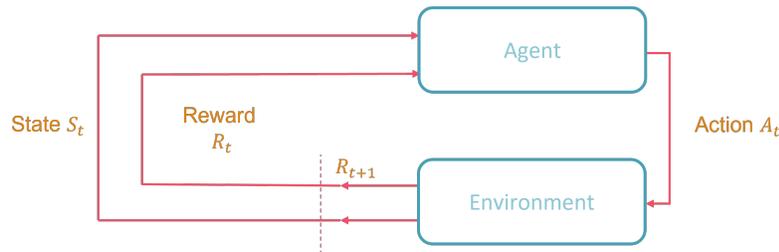


FIGURE 0.6 – Agent/Environment Interaction in the context of *reinforcement learning*

In the context of asset liabilities management (ALM), the RL agent can be considered as the portfolio manager who decides the allocations of each of the assets.

Among the reinforcement learning methods, the Deep Deterministic Policy Gradient (DDPG) is the algorithm selected and implemented in the ALM model within this study. The choice of this model is related to the fact that the algorithm is suited to optimization problems in continuous action spaces. The unique design of DDPG combines an Actor who generates actions from the current state of the environment, and the Critic who evaluates these actions by calculating the expected value of future rewards. This duality promotes both an efficient exploration and exploitation of the environment.

The use of the DDPG algorithm revolves around two distinct frameworks : a risk-neutral environment and a real-risk environment, each defined by its own optimization goals that rely on metrics adapted to their specific context. In the risk-neutral environment, the focus is on market risk and the present value of future profits (PVFP), aiming to comply with regulatory standards while improving the financial performance of the insurance company. In the real-risk environment, the objectives focus on the yield rate and latent wealth. This combination provides an interesting synergy between the liabilities and assets for the insurer.

Once the metrics are chosen in each model, the next step involves implementing the DDPG models in each risk universe. The goal of DDPG is to optimize these indicators, either by maximizing or minimizing them depending on their nature. Thus, the study proposes to investigate four models each defined by specific asset allocation strategies :

- The PVFP strategy (Maximize PVFP)
- The market risk strategy (Minimize market risk)
- The PVFP/Market risk strategy (Maximize PVFP and minimize market risk)
- The Yield rate/Latent Wealth strategy (Maximize Yield rate and maximize latent wealth)

Each DDPG model, after its training phase, determines an optimal asset allocation according to the defined metrics. These strategies optimized through the DDPG model are then compared to the existing strategy, based on a Fixed-mix approach. The methodology for optimizing the strategies adopted is summarized in the schema 0.7 presented below :

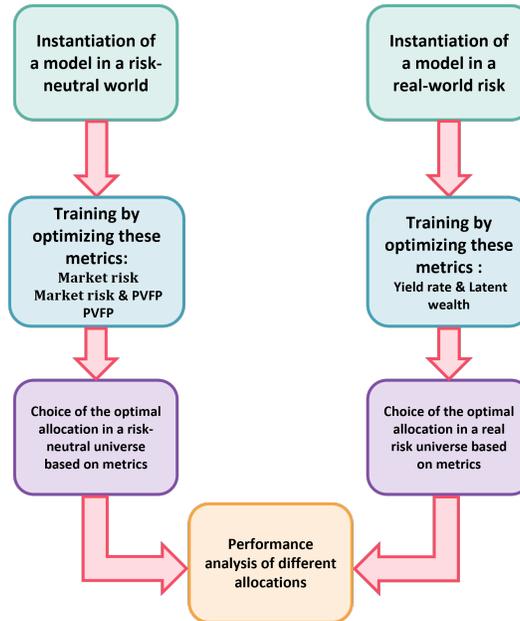


FIGURE 0.7 – Description of the methodology

The training phase of DDPG in the ALM model is a key step in developing an optimized asset allocation strategy. The agent’s state accounts for the metric values at the end of projection depending on the studied strategy (e.g., PVFP and market risk for the PVFP/Market risk strategy), including 40 values for each asset class (equity, real estate, and bonds) corresponding to the data extracted from the GSE and the 40 years of asset allocation projections.

The action space represents the possible allocation percentages among the different asset classes. Allocation constraints are also applied to ensure the feasibility and compliance of the strategies, limiting, for example, the investment proportions in each asset class and the percentage variation of an asset from one time step to another to avoid abrupt behaviors.

The reward function, as introduced in the schema 0.6, is designed to guide the agent towards optimal actions, ensuring that it adheres to all imposed constraints. During the calibration phase, the model was subjected to training of 200 episodes, each consisting of 1,000 individual simulations. This methodological approach results in a set of 200,000 learning trajectories. Additionally, the integration of a noise mechanism, specifically through the Ornstein-Uhlenbeck process, provides a systematic way to encourage exploration, thus avoiding the agent’s stagnation at local optima. This approach enhances the model’s robustness by increasing its generalization capacity through the testing of varied allocations. At the end of the 200,000 trajectories, the model observes which projection maximizes its reward. The selected trajectory then defines the asset allocation strategy to adopt. The optimal

allocations by assets, resulting from the simulations, are detailed below :

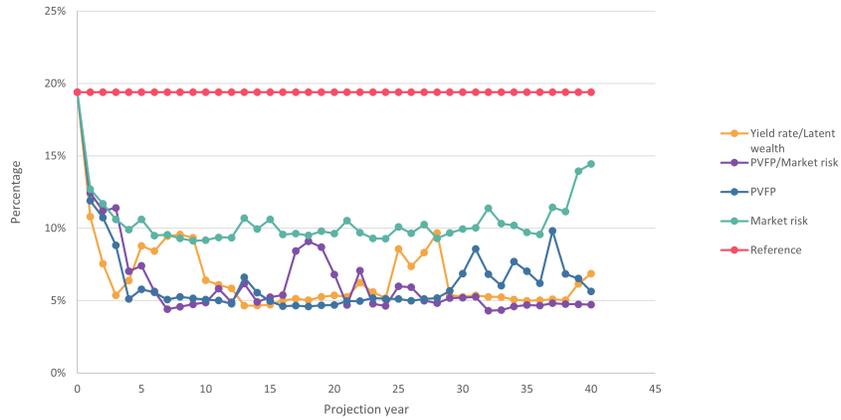


FIGURE 0.8 – Percentage of equity in the portfolio according to different asset allocation strategies

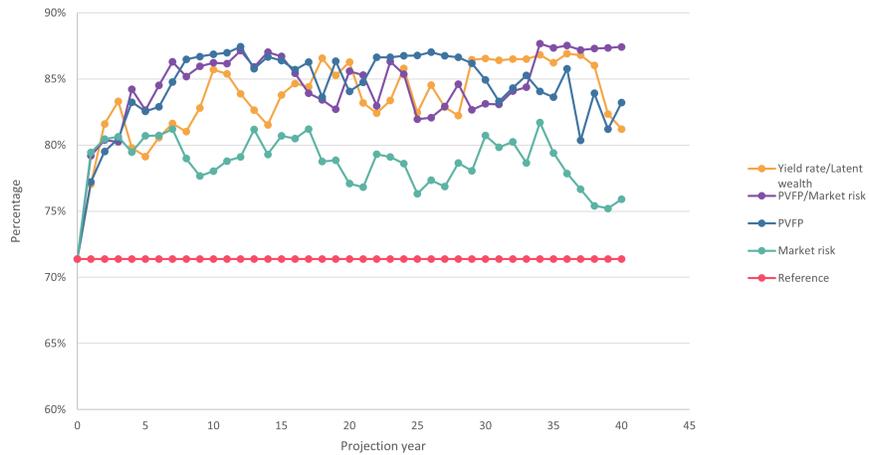


FIGURE 0.9 – Percentage of bonds in the portfolio according to different asset allocation strategies

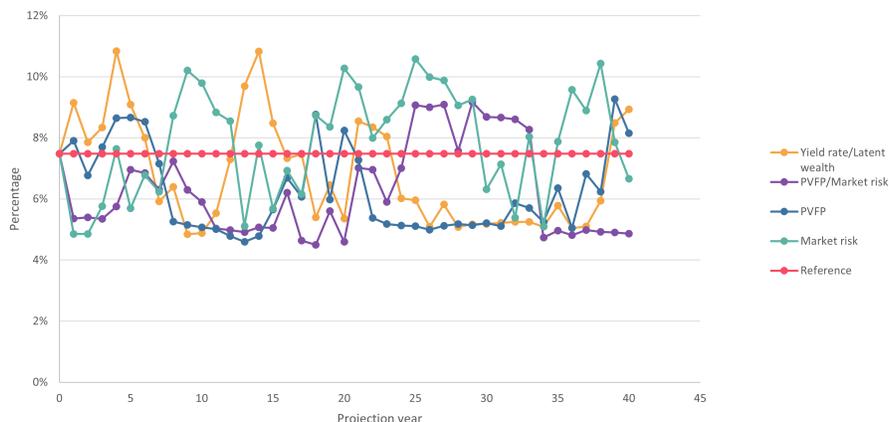


FIGURE 0.10 – Percentage of real estate in the portfolio according to different asset allocation strategies

The figures above (0.8, 0.9, and 0.10) illustrate the asset class distribution of optimal allocation strategies generated by DDPG. For each of the strategies, the model evaluates the metrics (PVFP, market risk, and SCR) in the risk-neutral environment, and the results are presented in the table 0.4 below. This approach aims to provide a point of comparison with the different strategies provided by DDPG and the reference Fixed-Mix strategy.

Strategy \ Metric	Reference	PVFP/Market risk	PVFP	Market risk
PVFP	3 449 110	3 857 009 ↑	3 885 879 ↑	3 765 291 ↑
SCR	2 672 117	2 614 645 ↓	2 660 763 ↓	2 569 768 ↓
Market risk	1 955 249	1 938 986 ↓	1 988 921 ↑	1 874 316 ↓

TABLE 0.4 – Summary of the strategies efficiency on different metrics compared to the reference scenario

Following the initial evaluation of strategies with training in a risk-neutral universe, the optimization strategy in a real-risk training universe consists of improving the yield rate and latent wealth. The obtained results show that the algorithm succeeded in significantly increasing the yield rate by massively selling latent wealth in equities and buying new bonds, thus generating a PVFP of more than 11 %. However, this strategy was accompanied by a decrease in latent wealth and a deterioration of SCR, making this approach less attractive in a regulatory context. To verify the model’s adaptability, a sensitivity analysis is conducted.

The sensitivity analysis aims to observe the model’s behavior by adjusting various parameters, to study their influence on the performance and stability of the system. It addresses two critical aspects to evaluate the robustness and adaptability of the DDPG model : sensitivities associated with

its calibration and sensitivities associated with economic conditions. The strategy optimizing the PVFP/Market risk pair is chosen to test these impacts.

For model sensitivities, three axes are studied : reward structure, the number of training episodes, and optimization of the PVFP/Market risk pair. For the sensitivity of the number of training episodes, the model is trained on a variety of episode numbers, ranging from 50 to 250. The objective is to detect the point where an increase in the number of episodes no longer results in a significant improvement in metrics, also considering the associated computational costs. The results of this thesis showed that the 200-episode model appears as the most optimal choice for the metric optimization/computational time pair.

The second sensitivity related to the model studied focuses on the reward. It consists of modifying the reward structure's calculation by introducing weights. This approach allows steering the optimization of one metric over the other variable of the study. For this, several weight combinations were tested :

- "Reference" with 0.5 weight for each metric
- "Favors PVFP" with 0.75 for PVFP and 0.25 for market risk
- "Favors Market risk" with 0.25 for PVFP and 0.75 for market risk

Strategy \ Metric	SCR	Market risk	PVFP
Reference	2 614 645	1 938 986	3 857 009
Favors PVFP	2 612 564 ↓	1 932 831 ↓	3 853 621 ↓
Favors Market risk	2 585 805 ↓	1 903 632 ↓	3 855 377 ↓

TABLE 0.5 – Results of sensitivity on reward structure

The table 0.5 above on the sensitivity of weights assigned in the reward function shows an influence on the optimization of market risk, with a reduction observed in the strategy favoring this criterion. However, the strategy focused on PVFP reveals only a slight decrease in the latter, suggesting a plateau in its optimization. This indicates that changing the weights can effectively improve Market risk, while the impact on PVFP remains limited. This possibly indicates that PVFP optimization has reached a plateau in this economic context. Thus, despite favoring PVFP optimization through weight adjustment, the model did not manage to achieve significant improvements in this metric.

The last model sensitivity analysis focuses on optimizing the PVFP/Life subscription risk pair. This sensitivity aims to study the impact of a metric more closely related to liabilities. However, it reveals a more modest efficacy. While this approach manages to improve life subscription risk by -1 %, the increase in PVFP is relatively limited, amounting to only +1 %. This result suggests that the model captures less effectively the interactions related to liabilities.

To deepen the analysis, the study of economic sensitivities focuses on two aspects : the impact of bond latent wealth by replicating the economic situation as of 31/12/2021 and the effects of the impact of the yield curve with a -100 bps shock.

Sensitivity \ Metric	SCR	Market risk	PVFP
Bonds latent wealth	0 %	-3,6 %	+7,9 %
Interest Rate Curve	-3,5 %	-5,1 %	+0,2 %

TABLE 0.6 – Results of economic sensitivities compared to the Fixed-Mix strategy

The introduction of bond latent wealth coupled with an economic environment as of 31/12/2022 corresponds to a favorable context for the insurer. Table 0.6 illustrates the model’s ability to exploit favorable market conditions to enhance overall performance. The results indicate that the model effectively captures latent wealth opportunities, testifying to its flexibility and adaptation to positive market dynamics.

Also, in Table 0.6, the sensitivity related to the change in the interest rate curve highlights a reverse effect on the metric favored by the model. Faced with a sudden drop in interest rates of -100bps, the model prioritizes market risk over PVFP. This result aligns with the new economic context of the shocked interest rate curve of -100bps. Indeed, the situation of lower interest rates complicates the release of financial results.

This work has explored the effectiveness of reinforcement learning, and more specifically of DDPG, in proposing an asset allocation strategy. The encouraging results of this study prompt further exploration of the potential of reinforcement learning. It’s important to note that, despite its advantages, the algorithm can lead to significant computation times, requiring special attention when deployed in production environments. This algorithm is constructed to provide the insurer with a versatile and agile method according to its performance or regulatory steering needs. DDPG emerges as a strategic tool for navigating the complex ecosystem of insurance, addressing the challenges posed by market fluctuations and evolving regulatory frameworks.

Introduction

L'activité assurantielle est marquée depuis 2022 par une rupture de la tendance baissière des taux. Le taux de rendement moyen des fonds euros, selon l'ACPR, s'est établi à 1,91 % en 2022 contre 1,28 % en 2021 enregistrant une hausse de 63 bps d'après les prévisions de l'ACPR pour 2023, le taux s'établit à 2,6 %. De plus, les augmentations successives du taux du Livret A, passant de 0,5 % en janvier 2022 à 3 % en février 2023, n'incitent pas les épargnants à opter pour des placements à moyen ou long terme comme l'assurance vie, entraînant une baisse de la collecte brute totale de -4,7 milliards d'euros par rapport à 2021. Dans un contexte d'inflation croissante, la priorité des assurés est donnée à l'épargne de précaution plus que jamais. L'environnement financier, marqué par des taux élevés affectant les contrats d'assurance vie, démontre la nécessité d'une gestion actif-passif (ALM) agile et innovante afin de répondre aux exigences réglementaires tout en restant compétitif envers les assurés en matière de rendements. En effet, les assureurs historiques sont confrontés à l'inertie de leur portefeuille obligataire construit en période de taux bas. Cette situation les handicape dans leur capacité à servir les taux de valorisation des fonds en euros aux niveaux actuels du marché. L'objectif de ce mémoire est de proposer une alternative à la stratégie d'allocation d'actifs actuelle de l'assureur permettant de répondre aux enjeux d'un contexte réglementaire et de taux élevés, le but étant de rester compétitif tout en gardant une gestion saine et optimale.

Pour répondre à cette problématique, une approche basée sur le *Deep Deterministic Policy Gradient* (DDPG), une méthode de *reinforcement learning*, est utilisée. Cette méthode a pour objectif de transformer la stratégie d'allocation statique du modèle ALM (*Fixed-Mix*) en une allocation dynamique. Elle a pour ambition d'optimiser l'allocation des actifs en restant en conformité avec les exigences réglementaires.

La première partie de ce mémoire établit l'environnement de travail du sujet. Elle introduit le contexte de la norme Solvabilité II et décrit en détail le modèle ALM, avec les mécanismes sous-jacents qui le composent. Les générateurs de scénarios économiques sont également abordés ainsi que leurs utilités dans les modèles ALM pour simuler différents environnements de marché. Les indicateurs de solvabilité et de rentabilité sont définis afin de pouvoir assurer aussi bien, le suivi, que la performance sur le plan économique du portefeuille.

Dans un second temps, un état de l'art est effectué en explorant les méthodes traditionnelles de gestion, permettant de dresser les avantages et inconvénients de ces méthodes. L'approche par *reinforcement learning* a pour but de *challenger* ces méthodes classiques. Pour cela, les différents concepts clés du *reinforcement learning* sont définis ainsi que le cadre théorique nécessaire à l'application de ces techniques pour l'ALM.

La troisième partie présente l'aspect pratique de cette étude avec la description et la mise en application de la méthode DDPG au sein du modèle ALM. Elle aborde les choix de modélisation et présente également le processus d'apprentissage du modèle en environnements de risque neutre et réel.

Enfin, une étude comparative est ensuite réalisée au travers de différents indicateurs permettant ainsi de comparer la stratégie existante et la méthode DDPG. Cette démarche permet d'aboutir à une analyse comportementale et de performance du modèle dans un contexte de taux élevés, déterminant la pertinence et l'efficacité de la démarche proposée.

Chapitre I

Introduction au contexte réglementaire et modélisation de l'ALM



L'assurance vie, de par sa complexité, est un produit financier qui nécessite une gestion prudente à la fois sur le court terme et sur le long terme. Les compagnies d'assurance, en raison de l'inversion du cycle de production (principe où l'assureur perçoit les primes avant d'effectuer d'éventuelles prestations), investissent les primes de leurs clients dans des actifs à long terme, notamment des obligations et des actions, en vue de générer des rendements suffisamment élevés pour couvrir leurs engagements futurs en matière de prestations. Cependant, depuis 2022, l'assurance vie a été confrontée à un environnement de croissance des taux d'intérêt comme l'illustre la figure I.1, ce qui a compliqué la tâche des assureurs historiques de générer des rendements suffisants pour maintenir leur rentabilité, leur solvabilité et leur compétitivité vis-à-vis de la concurrence. En effet, dans un environnement de taux d'intérêt élevés, ils se heurtent à l'inertie de leur portefeuille obligataire, constitué avec des obligations émises en période de taux bas. Cette situation entrave leur capacité à offrir des taux de valorisation sur les fonds en euros qui soient en phase avec les niveaux actuels du marché.

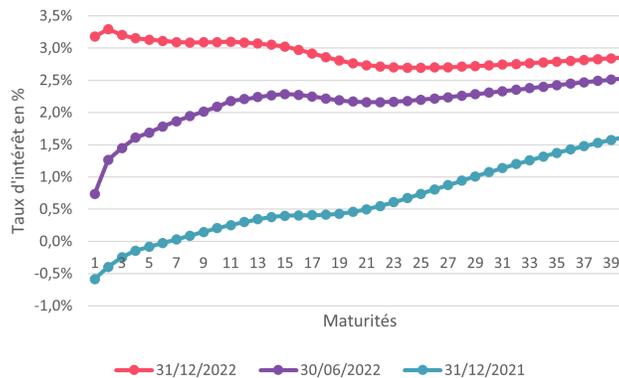


FIGURE I.1 – Courbes des taux EIOPA sans V.A

I.1 Spécificités de l'assurance vie

L'assurance vie constitue une catégorie spécifique au sein du secteur de l'assurance, caractérisée par des engagements liés à la vie humaine. L'assurance vie offre aux souscripteurs la possibilité d'épargner dans des actifs variés, allant de faiblement risqués à très risqués, dans le but de générer des intérêts et de transmettre ultérieurement cette épargne aux bénéficiaires, en cas de décès ou de survie de l'assuré. Les primes collectées sont investies sur les marchés financiers pour produire des rendements.

Par la suite, les différents supports d'épargne disponibles aujourd'hui en assurance vie seront détaillés.

- **Le fonds en euros** : Le fonds en euros représente une option d'épargne pour les investisseurs en quête de sécurité, permettant de placer leur capital à l'abri des fluctuations du marché. En effet, dans ces contrats en euros, l'assureur s'engage à préserver en permanence le capital initial de l'investisseur. Principalement investi en obligations d'État (OATs) et obligations

d'entreprises (corporate), avec une portion résiduelle allouée aux actions, à l'immobilier ou aux produits dérivés. Quatre critères essentiels définissent les contrats d'épargne en euros. Le **taux technique** dans un contrat d'assurance vie est le taux minimal de revalorisation garanti sur la durée du contrat, fixé dès la souscription. Ce taux, soumis à un plafond réglementaire, assure aux souscripteurs une valorisation minimale de leur épargne, indépendamment des fluctuations des marchés financiers. En complément, l'assureur peut également offrir un **Taux Minimum Garanti (TMG)**, similaire au taux technique, mais applicable uniquement sur une période maximale de deux ans. Enfin, la **participation aux bénéficiaires (PB)** mécanisme qui sera détaillé dans la partie I.4.2 et le TMG permettent ainsi de calculer le **taux servi** à l'assuré.

- **Le fonds en unités de compte (UC)** : Le risque est porté par l'assuré, l'assureur garantissant un nombre de parts d'unités de compte plutôt qu'un montant fixe en euros. La valeur des parts dépend du support sur lequel elles sont investies et fluctue donc en fonction des marchés financiers. Ces fonds, généralement plus volatiles que les fonds en euros, peuvent entraîner des plus-values ou moins-values significatives, dues aux aléas des marchés boursiers.

Les deux familles de contrats offrent une garantie commune : la garantie de rachat. Elle permet aux assurés de récupérer, à tout moment, l'épargne accumulée depuis le début du contrat.

Les contrats d'assurance vie se répartissent en deux types de familles. D'une part, les contrats **monosupports** qui permettent à l'épargnant d'investir exclusivement dans un seul type de fonds, soit en euros, soit en unités de compte, offrant ainsi une approche ciblée selon les préférences de l'investisseur. D'autre part, il existe les contrats **multisupports**, caractérisés par la combinaison d'au moins un fonds en euros et de plusieurs supports en unités de compte, ce qui permet une distribution des primes versées par le souscripteur entre divers supports alignés sur son profil de risque, favorisant une stratégie d'investissement diversifiée.

Dans le cadre de cette étude, le portefeuille se compose de contrats multisupports avec un capital réparti approximativement à 20 % en UC et à 80 % en fonds euros. Les caractéristiques du portefeuille seront détaillées dans la partie I.6.

Après avoir présenté l'assurance vie d'un point de vue global, il est essentiel d'introduire le cadre réglementaire de cette étude : la norme Solvabilité II.

I.2 La réglementation Solvabilité II

La réglementation Solvabilité II (S2) désigne un cadre réglementaire fixant le régime de solvabilité des compagnies d'assurances, de réassurances et des mutuelles au sein de l'Union Européenne (UE). Son but est d'assurer une harmonisation des règles au sein de l'UE en matière de capital afin que les compagnies puissent honorer leurs engagements envers les assurés et bénéficiaires. Ce régime se

traduit notamment par un niveau d'exigence en capital, mais aussi en matière de gouvernance et de *reporting*.

Les objectifs majeurs de Solvabilité 2 sont les suivants :

- Consolider l'harmonisation du marché européen de l'assurance.
- Optimiser la protection des souscripteurs d'assurance et des bénéficiaires.
- Accroître la compétitivité des compagnies d'assurance et de réassurance européennes sur la scène internationale.
- Mettre en place une approche par évaluation économique du bilan.
- Appréhender la volatilité des sinistres afin de mieux évaluer leurs coûts réels

Les trois piliers de Solvabilité II

Solvabilité II repose sur trois piliers interdépendants comme illustrés de manière simplifiée par la figure I.2. Ces trois piliers couvrent l'ensemble des aspects, de la quantification des risques à la gouvernance, en passant par la transparence. Cette étude se concentre spécifiquement sur le pilier 1 de la norme.

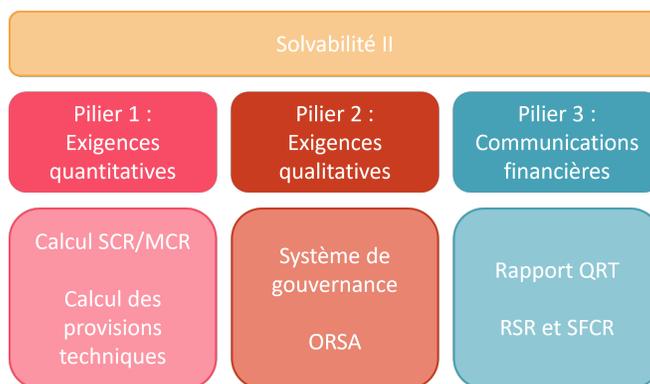


FIGURE I.2 – Piliers de la réforme Solvabilité II

Focus sur le pilier 1 : Exigences quantitatives

Le pilier 1 de Solvabilité II porte sur les exigences de fonds propres que les assureurs et réassureurs doivent détenir pour se prémunir contre les risques inhérents à leurs opérations. Il détaille la méthodologie de calcul du capital réglementaire ainsi que l'évaluation des actifs et des passifs. Solvabilité II impose une valorisation des actifs et des passifs en valeur de marché. Les provisions techniques sont

définies comme la combinaison du BE¹ (*Best Estimate*), représentant la meilleure estimation des flux futurs, et de la RM² (*Risk Margin*), qui est la marge de risque associé à l'incertitude du BE. Un bilan sous Solvabilité II est visualisé sous la forme suivante I.3 :

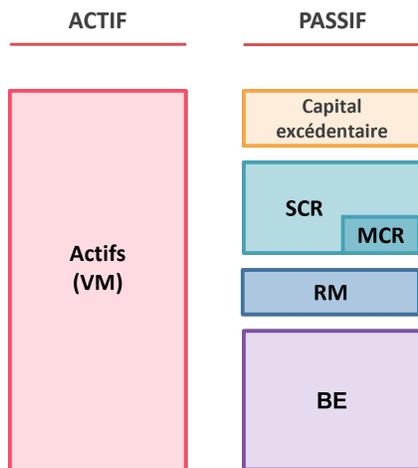


FIGURE I.3 – Bilan Solvabilité II

Le SCR par la formule standard

Le SCR (*Solvency Capital Requirement*) est le pilier de la réforme Solvabilité II. Il représente le montant de capital que l'assureur ou le réassureur doit détenir pour couvrir les risques auxquels l'entreprise pourrait être confrontée dans l'année à venir, avec une probabilité de 99,5 %. En d'autres termes, il s'agit du capital nécessaire pour que la compagnie d'assurance puisse faire face à des scénarios de perte extrême.

Il y a principalement deux méthodes pour calculer le SCR :

- **Les modèles internes** : Avec l'approbation des régulateurs, les entreprises peuvent développer leurs propres modèles statistiques et économétriques pour évaluer leur SCR, en fonction de leur propre profil de risque.
- **La formule standard** : Il s'agit d'une formule prédéfinie fournie par les régulateurs qui prend en compte divers sous-modules de risques tels que le risque de marché, le risque de souscription en assurance vie et non-vie, le risque de contrepartie, etc. La formule standard sera privilégiée dans ce mémoire puisqu'elle est la solution la plus représentative de la globalité de marché.

Afin d'obtenir le SCR, la formule doit prendre en compte plusieurs modules de risques auxquels sont exposés les assureurs. La cartographie des risques ci-dessous I.4 communiquée par l'EIOPA permet

1. la notion de *Best Estimate* est expliquée dans la partie I.7
2. la notion de *Risk Margin* est expliquée dans la partie I.7

de représenter les différents modules de risques. L'assureur est confronté dans le cadre de l'assurance vie à deux risques principaux que sont le risque de marché et le risque vie. Le SCR des différents sous-modules est calculé selon la méthodologie suivante : un choc de stress est appliqué sur le scénario central afin de mesurer l'impact sur le bilan.

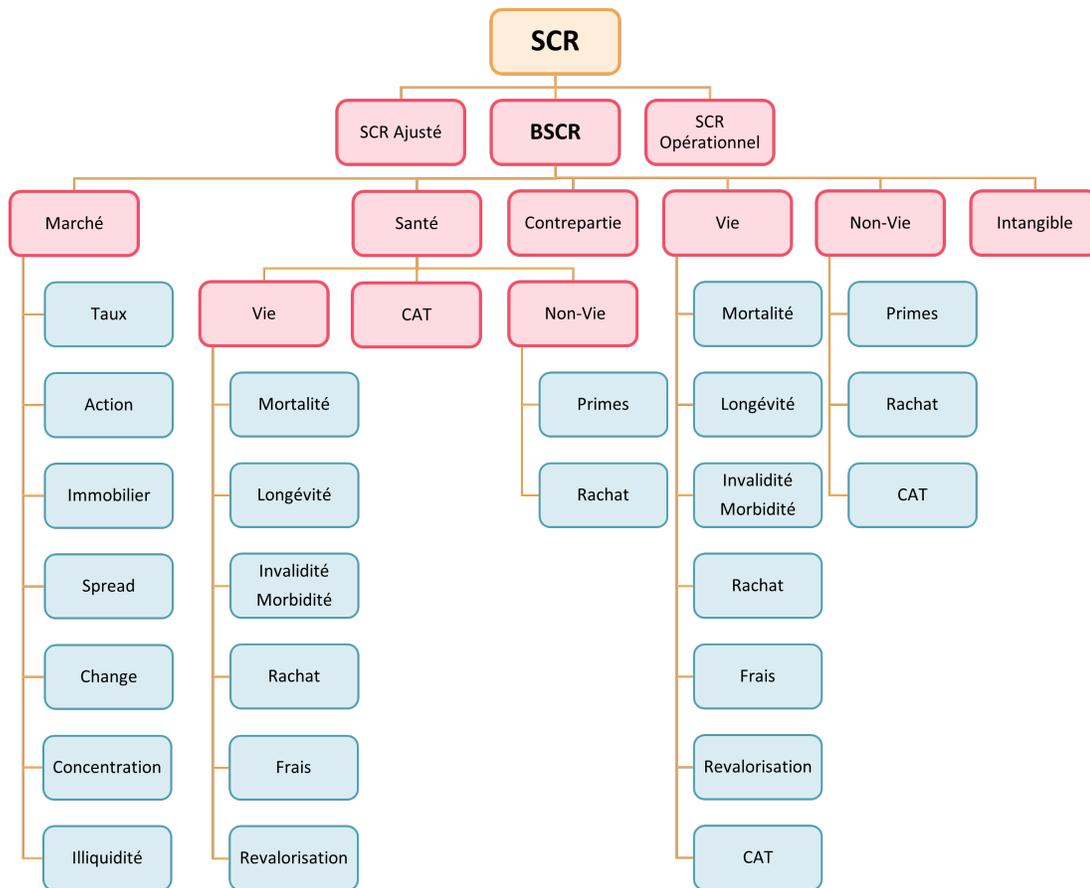


FIGURE I.4 – Pieuvre du SCR

Afin de déterminer le SCR en suivant la méthodologie établie précédemment, l'utilisation de la NAV (*Net Asset Value*), soit la valeur de l'actif net, est nécessaire. La NAV dans un scénario central est comparée avec la NAV dans un scénario choqué comme l'illustre la figure I.5. Un scénario choqué fait référence au contexte où le risque étudié provoque une situation défavorable et dont l'assureur doit détenir un montant minimum de capital afin de limiter la ruine à une probabilité de 5 %. Considérons un sous-module de risque donné x dont le SCR est calculé à partir de la différence entre la NAV du scénario central et la NAV choquée, décrit dans l'équation suivante :

$$SCR_x = NAV_{central} - NAV_{choc} \quad (I.1)$$

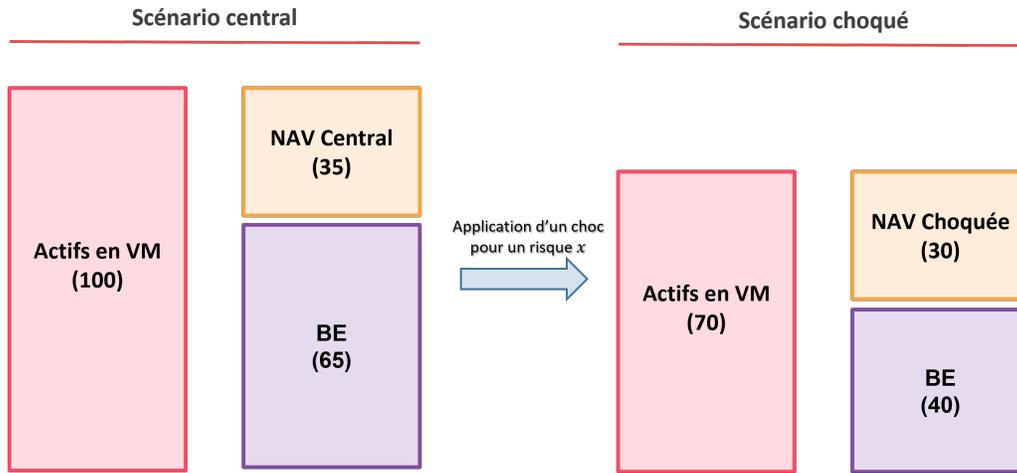


FIGURE I.5 – Calcul du SCR pour un risque donné

Dans l'exemple ci-dessus, la valeur du SCR_x est égale à 5. Après avoir déterminé tous les sous-modules SCR_i , où $i \in \{\text{actions, spread, ...}\}$, le $SCR_{\text{marché}}$ (I.2) est évalué en agrégeant les SCR_i à l'aide d'une matrice de corrélation fournie par le régulateur, disponible en Annexe A.1. Cet indicateur de solvabilité sera employé dans la suite de l'étude pour évaluer la pertinence de la stratégie d'allocation d'actifs.

$$SCR_{\text{marché}} = \sqrt{\sum_{(i,j) \in n \times n} \rho_{ij} \times SCR_i \times SCR_j} \quad (I.2)$$

avec :

- n le nombre de sous-modules présents dans le module Marché.
- ρ_{ij} le coefficient de corrélation entre les sous-modules de risques i et j .

En suivant la méthode de la formule standard, vient ensuite le calcul du BSCR (*Basic Solvability Capital Requirement*) (I.3) dont découlera par la suite le SCR (I.4). Les formules respectives sont rappelées ci-dessous. La méthodologie de calcul de ces deux éléments est expliquée de façon plus détaillée dans le mémoire d'actuariat de M.MASSON [2021].

$$BSCR = \sqrt{\sum_{(i,j) \in m \times m} \rho_{i,j} \times SCR_i \times SCR_j} + SCR_{\text{intangible}} \cdot \quad (I.3)$$

où :

- m le nombre de modules de risque principaux.
- $SCR_{\text{Intangible}}$ correspond à 80% des actifs incorporels qui peuvent être des brevets ou de la technologie. Dans le modèle ALM de cette étude, ce SCR ne sera pas utilisé et sera fixé à une valeur de 0.

Enfin, le BSCR permet de calculer à l'aide de deux autres composantes le SCR final dont la formule est la suivante :

$$\text{SCR} = \text{BSCR} - \text{Adj} + \text{SCR}_{op}. \quad (\text{I.4})$$

avec :

- Adj est l'ajustement au titre des impôts différés et des pertes des provisions techniques.
- SCR_{op} est le risque associé au personnel, la défaillance des logiciels ou du matériel utilisé.

Après avoir déterminé le SCR, qui représente donc le capital nécessaire pour faire face aux scénarios de risques extrêmes, il est essentiel que l'assureur adopte une position par rapport à cette exigence. Le ratio de solvabilité est un indicateur qui permet de mesurer cette position en utilisant les fonds propres éligibles au SCR :

$$\text{Ratio de solvabilité} = \frac{\text{Fonds propres éligibles au SCR}}{\text{SCR}} \quad (\text{I.5})$$

Un ratio supérieur à 100 % indique que l'assureur possède plus de capital que le montant requis par le régulateur. En réalité, les régulateurs incitent les assureurs à avoir un coussin de solvabilité au-dessus du SCR. Le modèle ALM (*Assets Liabilities Management*) ne calculant pas les fonds propres, il a été décidé de ne pas considérer le ratio S2, mais un indicateur de solvabilité en prenant en compte la valeur in force³ (VIF) :

$$\text{Indicateur de solvabilité} = \frac{\text{VIF}}{\text{SCR}} \quad (\text{I.6})$$

Afin d'appréhender ces risques, il est primordial pour l'assureur de procéder à une bonne gestion actif-passif afin de minimiser les différents facteurs de risques pouvant impacter l'actif et/ou le passif. La gestion actif-passif ou encore ALM permet de répondre à cette problématique.

3. La définition de la VIF est fournie dans la partie I.7

I.3 Définitions et enjeux de l'ALM

La gestion actif-passif est le processus continu de formulation, de mise en œuvre, de suivi et de révision des stratégies liées à l'actif et au passif afin d'atteindre les objectifs financiers, pour un ensemble donné de tolérances et de contraintes en matière de risque.

L'assureur investit un capital sur différents actifs afin que les flux financiers découlant de celui-ci permettent de répondre aux flux probables de passif, tout en optimisant son investissement. Cependant, il ne faut pas oublier que ce dernier doit aussi garder un certain niveau de liquidité pour pouvoir répondre à tout instant aux engagements pris envers les assurés. Il est donc important lors de l'élaboration d'une stratégie ALM de connaître les différents produits du passif pour anticiper les évolutions dans le but de réduire les risques liés à l'activité d'assurance. La dynamique ALM se modélise alors par le schéma I.6. Une somme initiale provenant des primes assurées est investie dans un ensemble d'actifs, ces actifs produisent des rendements qui sont ensuite ajoutés au montant initial. Ce montant permet par la suite de faire face aux différents engagements de l'assureur sur l'année qui sont représentés en rouge sur la figure. Ce processus est ensuite répété un nombre de fois suffisamment jusqu'à ce que l'assureur n'ait plus aucun engagement à respecter auprès de ses assurés.

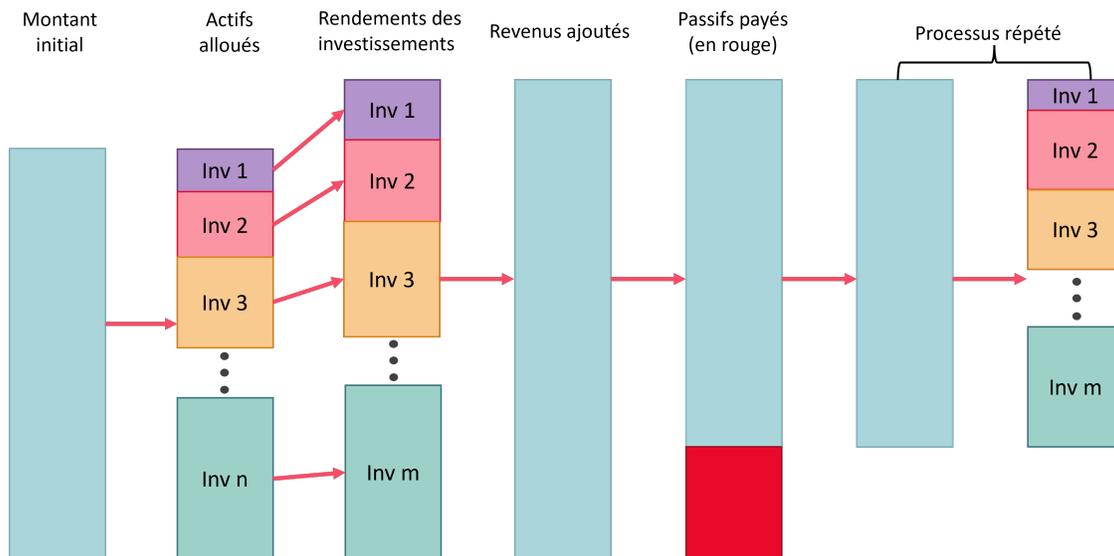


FIGURE I.6 – La dynamique ALM (Rodrigues Fontoura, [2020])

Par conséquent, l'ALM dans un contexte économique concurrentiel consiste à placer de manière optimale son investissement, selon une politique de pilotage tout en satisfaisant ces engagements que se soit en matière de rendement ou normatif dans le cadre Solvabilité II . La section suivante présente en détail le modèle ALM étudié avec ses mécanismes sous-jacents.

I.4 Présentation du modèle ALM utilisé

I.4.1 Fonctionnement du modèle ALM

Le modèle ALM utilisé dans le cadre de cette étude est un modèle ALM d'Optimind implémenté à l'aide du langage de programmation VBA. Ce modèle est constitué de plusieurs modules interdépendants. Pour fonctionner, le modèle a besoin de plusieurs jeux de données. Une partie des données à l'actif proviennent d'un module indépendant du modèle ALM qui a pour but de réaliser des projections de grandeurs économiques : le GSE⁴. Une fois les données entrées dans le modèle, elles vont subir différentes étapes illustrées à l'aide de la figure I.7 :

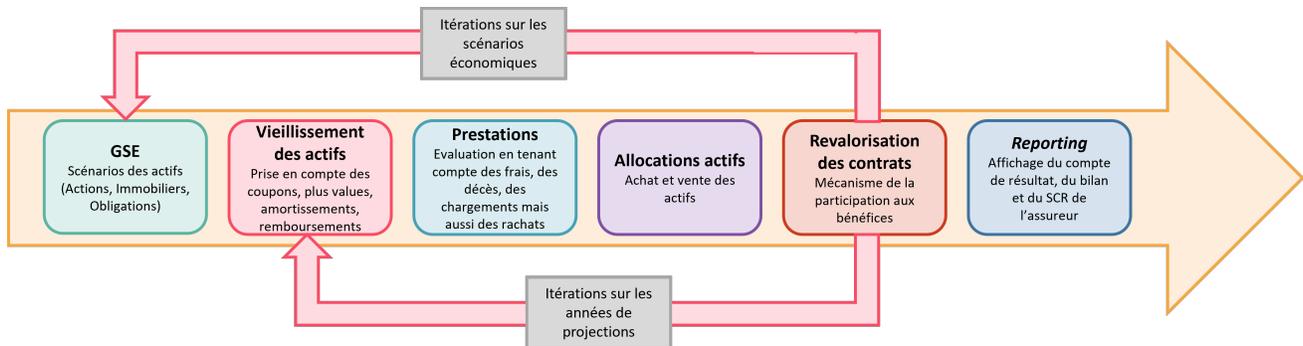


FIGURE I.7 – Fonctionnement du modèle ALM

Les différents indicateurs sont obtenus à l'aide de la méthode de Monte-Carlo en effectuant la moyenne des valeurs obtenues sur les différents scénarios économiques. Dans le cadre de cette étude, le modèle réalise des projections sur un horizon de 40 ans avec un pas annuel. Le nombre de simulations dépend du besoin associé. Il existe deux types de scénarios de simulation :

1. Scénario déterministe (ou équivalent-certain) :

- Ce scénario est établi sur des hypothèses pré-déterminées, sans tenir compte de la variabilité ou de l'incertitude. Le scénario permet de bien développer le modèle ALM et de s'assurer de la validité de celui-ci. En effet, l'écart entre ce qui est attendu au passif et à l'actif doit être nul. Cette différence est appelée écart de convergence I.7. Il permet donc de déceler d'éventuelles erreurs de modélisation.
- Le scénario est employé pour évaluer le coût des garanties associées aux contrats, tels que le taux minimum garanti (TMG), la participation aux bénéfices (PB), l'option de rachat, ou l'arbitrage. Ce coût est déterminé en observant l'écart entre les fonds propres obtenus du scénario déterministe et ceux provenant du scénario stochastique central.

4. La notion de GSE sera détaillée dans la partie I.5 de ce mémoire

2. Scénario stochastique :

Dans le cas des simulations stochastiques, l'écart de convergence doit tendre vers 0 au fur et à mesure que l'on augmente le nombre de simulations. Dans le cadre de cette étude, le nombre de simulations est égal à 1000, l'ordre de grandeur de l'écart de convergence attendu pour le modèle est de 0,05 %. Cela permet d'évaluer avec robustesse les indicateurs réglementaires.

L'écart de convergence est calculé depuis le bilan Solvabilité 2 à l'aide de la formule ci-dessous :

$$\text{écart de convergence} = \frac{\text{total passif} - \text{total actif}}{\text{total actif}}. \quad (\text{I.7})$$

I.4.2 Description des mécanismes du modèle

Le modèle ALM, tel qu'introduit dans la partie précédente, est un agrégat de différentes étapes qui permettent de retranscrire le processus des contrats d'assurance vie. Pour cela, il est nécessaire d'introduire les différentes méthodes de calcul qui influent sur ces contrats.

Politique de taux servis et participation aux bénéfices :

Les contrats euros sont constitués d'une garantie que l'assureur doit servir au minimum : le taux minimum garanti (TMG). De plus, la loi (Article L.331-3 [2023]) impose à l'assureur de redistribuer une partie de ses résultats techniques et financiers. Cette disposition s'inscrit dans le cadre de la Participation aux Bénéfices (PB), un mécanisme clé en assurance vie qui assure que les assurés bénéficient de manière discrétionnaire des performances financières de l'assureur.

La PB réglementaire est donnée par :

$$PB_{\text{réglementaire}} = 0.85 \times \text{Résultat}_{\text{financier}} + 0.90 \times \text{Résultat}_{\text{technique}} \quad (\text{I.8})$$

La politique de participation aux bénéfices est un mécanisme central des contrats ayant une partie en euros qui vise à rémunérer les assurés afin de permettre de fidéliser le portefeuille. Ce mécanisme est conçu pour permettre d'équilibrer les intérêts entre assureur et assurés, tout en respectant les exigences réglementaires et contractuelles. L'équation I.9 désigne le montant servi aux assurés comme l'exige le contrat. Néanmoins, afin de rester compétitif sur le marché l'assureur peut choisir un montant plus élevé, un montant cible.

$$PB_{\text{contractuelle}}(mp) = \text{Produits}_{\text{financiers}}(mp) \times \text{Taux}_{\text{contractuel}} - \text{Chgt} - \text{IT} \quad (\text{I.9})$$

où $\text{Taux}_{\text{contractuel}}$ est le taux de participation aux bénéfices établi contractuellement, Chgt les charges sur encours, IT les intérêts techniques, et $\text{Produits}_{\text{financiers}}$ la part des produits financiers⁵

5. La méthodologie de calcul des produits financiers est détaillée dans la partie I.7 de ce mémoire

attribués aux *model points*⁶ (MP).

Pour autant, la réglementation permet à l'assureur de ne pas distribuer instantanément l'intégralité de ces bénéfices. Il peut aussi placer une partie de la PB dans une provision : la provision de participation aux bénéfices (PPB). Cependant, l'assureur a l'obligation de redistribuer cette provision aux assurés dans un délai de 8 ans.

La PPB permet de lisser les rendements futurs des contrats d'assurance. Le montant versé est la somme du taux du TMG, de la PB immédiatement payée, des éventuelles reprises de la PPB et de la marge financière en cas d'ultime recours. Une limite du modèle actuel est que la PPB initiale est redistribuée de manière stricte (au lieu d'être lissée sur les 8 ans) à la 8ème année de projection entraînant un pic de taux servi.

Dans le cadre de ce modèle, trois scénarios de politiques de distribution de PB sont à envisager :

- La PB contractuelle est servie ;
- La PB contractuelle et le puisement dans la PPB afin de pouvoir servir un montant cible ;
- Une fois que la PB contractuelle et la PPB sont calculées, si la PPB ne suffit pas à atteindre le montant cible, une déduction de la marge financière devra s'ajouter pour atteindre la cible.

New Business :

Le *New Business* est un mécanisme du modèle permettant d'intégrer des nouveaux contrats agrégés en *model points* dans le portefeuille. Cela permet, dans une projection en monde réel de refléter la croissance attendue par rapport aux conditions du marché.

Dans le cas spécifique de ce modèle ALM, 81 *model points* de passif sont ajoutés chaque année pendant les 5 premières années de projections. Par la suite, le portefeuille est déroulé en *run-off*⁷. Ce mécanisme est décrit ci-dessous :

$$Model\ Points_t = \begin{cases} Model\ Points_{t-1} + 81 & \text{pour } t = 1, 2, 3, 4, 5 \\ Model\ Points_{t-1} & \text{pour } t > 5 \end{cases}$$

où $Model\ Points_t$ représente le nombre total de *model points* de passif à l'année t .

6. La définition d'un *model point* est fournie dans la partie I.6

7. désigne un ensemble de contrats d'assurance que la compagnie a cessé d'émettre ou de renouveler, mais qui restent en vigueur jusqu'à leur expiration naturelle ou leur liquidation.

Modélisation des rachats et des décès :

La modélisation des rachats est un aspect fondamental de la gestion des risques en assurance vie. Elle permet de comprendre et de prévoir le comportement des assurés sur leur capacité à racheter leur contrats. Les rachats sont à distinguer en deux catégories distinctes : **les rachats structurels** et **les rachats conjoncturels**.

Les rachats structurels sont liés à la fiscalité et aux facteurs exogènes aux rendements des contrats d'assurances. Le taux de rachats structurels, tel que modélisé dans le modèle ALM, augmente avec la durée de détention du contrat et se stabilise après la 8^{ème} année. Ce taux des rachats dépend notamment de la fiscalité avantageuse après 8 années de détention.

Les rachats conjoncturels sont sensibles aux conditions économiques. Ils sont déclenchés lorsque le taux servi est insatisfaisant, ce qui implique que les assurés rachètent leurs contrats afin de pouvoir bénéficier de rendements plus attractifs. Ces rachats sont modélisés par un seuil, avec un taux cible annuel, donné par l'OAT 10 ans. Si le taux cible est servi alors les assurés resteront dans le portefeuille.

Dans le cadre de cette étude, les rachats partiels ne sont pas modélisés. Les rachats totaux quant à eux sont calculés par la somme des rachats structurels et rachats conjoncturels.

De plus, la modélisation des décès joue un rôle crucial dans la gestion des contrats d'assurance vie. La provision mathématique est diminuée par les prestations versées en cas de décès. Dans ce contexte, il est supposé que la population du portefeuille est exclusivement masculine, et la table TH 00-02 est utilisée pour déterminer les taux de mortalité adoptant ainsi une approche plus prudente. Cette décision repose sur le fait que les femmes présentent une espérance de vie plus longue, or le risque de longévité est un élément clé en assurance vie.

Après avoir décrit le fonctionnement du modèle ALM, il est essentiel de comprendre la provenance des divers jeux de données en entrée de ce modèle. C'est ici que les GSE entrent en jeu.

I.5 Les générateurs de scénarios économiques

Un générateur de scénarios économiques (GSE) est un outil qui permet de simuler différentes trajectoires de variables économiques dans le futur. Cet outil permet en outre de simuler des taux, des valeurs d'actions, des valeurs d'immobilier et, de façon plus générale, toute grandeur économique. Les variations temporelles de ces indicateurs économiques sont générées à l'aide de modèles spécifiques pour chacun d'entre eux et permettent ainsi d'obtenir une partie des *inputs* du modèle ALM. Pour obtenir des valeurs pertinentes, il est nécessaire de calibrer au préalable les paramètres de ces modèles à l'aide de données de marché comme illustré par la figure I.8.

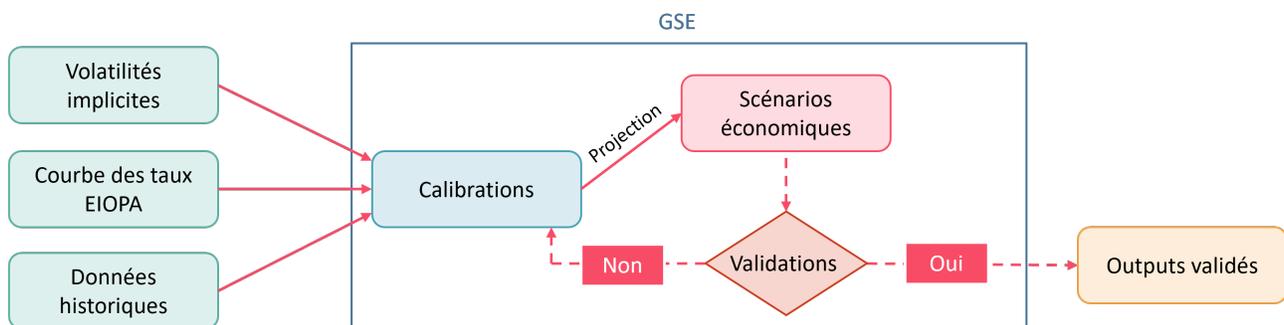


FIGURE I.8 – Construction et fonctionnement d'un GSE

Il existe deux cadres de projections :

- Le monde risque neutre pour une valorisation prudente.
- Le monde réel pour une perspective axée sur le risque.

I.5.1 GSE risque neutre

Dans un monde risque neutre, le rendement des actifs est en moyenne le taux sans risque. Cet univers repose notamment sur deux hypothèses clés : l'absence d'opportunités d'arbitrage et la complétude des marchés financiers, ce qui équivaut à l'existence d'une mesure de probabilité. La mesure risque neutre, généralement notée \mathbb{Q} , sous laquelle la valeur actualisée des actifs se comporte comme une martingale. Des tests doivent être réalisés afin de s'assurer que les projections obtenues respectent les hypothèses suivantes : cohérence avec le marché (les prix des actifs calculés sous la mesure \mathbb{Q} doivent être en accord avec les prix observés sur le marché) et test de martingalité. Dans le cadre de ce mémoire, on considère ces tests comme étant réalisés afin de valider ces hypothèses.

Courbe des taux :

Les scénarios économiques du GSE risque neutre sont générés à l'aide de la courbe des taux sans risque du 31/12/2022. La figure I.9 de la courbe des taux présente des taux élevés sur un horizon projeté de 40 ans, malgré une légère décroissance à partir de la maturité 15 ans. Cette courbe s'inscrit donc de manière pertinente dans le contexte actuel de taux élevés.

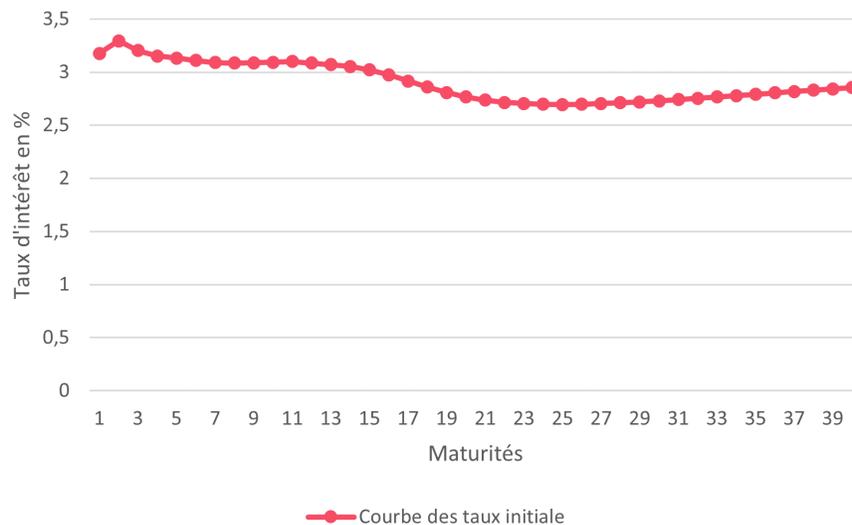


FIGURE I.9 – Courbe des taux sans risque fournie par l'EIOPA au 31/12/2022

Modèle de Hull & White :

Le modèle de Hull & White est utilisé pour projeter le taux d'intérêt. Il s'agit d'une généralisation du modèle de Vasicek. Il permet d'intégrer la courbe des taux initiale grâce au terme $\theta(t)$. L'équation différentielle stochastique du modèle Hull-White est donnée par :

$$dr_t = (\theta(t) - ar_t)dt + \sigma dW_t^{\mathbb{Q}} \quad (\text{I.10})$$

où :

- r_t : taux d'intérêt à court terme à l'instant t .
- $\theta(t)$: fonction du temps contrôlant la tendance de r_t .
- a : coefficient positif déterminant la vitesse de retour à la moyenne du taux.
- σ : volatilité du taux d'intérêt.
- $W_t^{\mathbb{Q}}$: un mouvement brownien standard sous la mesure de probabilité risque neutre.

Pour résoudre l'équation différentielle stochastique (EDS) du modèle Hull-White, une approche par discrétisation est utilisée afin de trouver une solution à l'équation. En utilisant le lemme d'Itô et en cherchant une solution sous forme d'exponentielles, on peut trouver une solution pour r_{t+h} (où h désigne un petit intervalle de temps) en fonction de r_t et d'autres paramètres.

La solution discrète exacte pour l'EDS du modèle Hull-White, similaire au modèle de Vasicek, est :

$$r_{t+h} = r_t e^{-ah} + \theta(t+h) - \theta(t)e^{-ah} + \sigma \sqrt{\frac{1 - e^{-2ah}}{2a}} Z \quad (\text{I.11})$$

où Z est une variable aléatoire suivant une distribution normale tel que $Z \sim \mathcal{N}(0,1)$ et θ est une fonction dépendant des prix des zéro-coupons à la date $t=0$.

Modèle de Black & Scholes :

Le prix des actions et de l'immobilier est modélisé au travers du modèle de Black & Scholes. Dans le contexte d'un univers risque-neutre, la dynamique du prix est la suivante :

$$dS_t = rS_t dt + \sigma S_t dW_t \quad \text{où } S_t \text{ est le prix de l'actif,} \quad (\text{I.12})$$

La solution exponentielle pour le prix est donnée par :

$$S_t = S_0 \exp \left(\int_0^t \left(r - \frac{\sigma^2}{2} \right) ds + \int_0^t \sigma dW_s \right)$$

Pour discrétiser cette équation sur un intervalle de temps h :

$$S_{t+h} = S_t \exp \left(\left(r - \frac{\sigma^2}{2} \right) h + \sigma \sqrt{h} Z \right), \quad (\text{I.13})$$

Ces deux modèles permettent de répondre aux exigences réglementaires de Solvabilité II, qui imposent des calculs dans un univers à risque neutre pour déterminer notamment le SCR, la PVFP, et d'autres indicateurs.

Après avoir introduit le GSE risque neutre, il est pertinent de se tourner vers le GSE en monde réel. En effet, cette approche permettra d'évaluer les performances de l'optimisation de l'allocation d'actifs de manière plus fidèle par rapport aux conditions de marché.

I.5.2 GSE risque réel

Le monde réel se caractérise par des rendements basés sur les attentes réelles des agents économiques et les tendances historiques des marchés. Dans un contexte d'étude d'une optimisation d'allocation

d'actifs, le monde réel est plus adapté afin de pouvoir utiliser des indicateurs de rentabilité comme la richesse latente ou encore le taux de rendement actuariel⁸. Par la suite, la probabilité du monde réel est notée \mathbb{P} .

Le modèle de Vasicek est utilisé pour les taux d'intérêt tandis que pour la valeur des actions et de l'immobilier, le modèle de Black & Scholes est employé comme introduit précédemment.

Modèle de Vasicek :

Le modèle de Vasicek est souvent utilisé pour décrire l'évolution des taux d'intérêt. Il s'agit d'une équation différentielle stochastique (EDS) de la forme :

$$dr_t = a(\theta - r_t)dt + \sigma dW_t \quad (\text{I.14})$$

où :

- a la vitesse de retour à la moyenne.
- θ la moyenne à long terme du taux d'intérêt.
- σ la volatilité du taux d'intérêt.
- W_t un mouvement brownien standard sous la mesure de probabilité risque réel.

En résolvant l'équation précédente I.14, il est obtenu en utilisant le lemme d'Itô :

$$r_t = e^{-at} \left(r_0 + \int_0^t a\theta e^{au} du + \sigma \int_0^t e^{au} dW_u \right) \quad (\text{I.15})$$

Pour calibrer ces modèles, les données choisies en entrée sont les suivantes : les cotations de l'IEIF pour l'immobilier, Euribor 3 mois pour les taux courts, l'indice des OAT 10 ans pour les taux longs et enfin l'EuroStoxx pour les actions.

Les sorties des GSE sont essentielles pour le bon fonctionnement du modèle ALM. Cependant, il est tout aussi important d'étudier l'état initial du portefeuille à l'actif et au passif afin de bien comprendre le fonctionnement du modèle ALM.

8. Ces métriques seront détaillées dans la partie I.7

I.6 Caractéristiques du portefeuille étudié

Le portefeuille étudié est fictif, représentant un assureur vie moyen au 31/12/2022 ajusté de manière à avoir une vision prudente (notamment le fait de prendre en considération que des assurés de sexe masculin). Des *model points* ont été établis tant pour l'actif que pour le passif. Un MP est un agrégat de contrats ou d'actifs ayant des caractéristiques similaires.

I.6.1 Le passif

Le passif du portefeuille du modèle s'articule autour de 81 *model points* représentant 81 000 polices d'assurance. Chaque MP étudié comprend donc un ensemble de 1000 polices. Les caractéristiques ainsi que les valeurs moyennes des *model points* sont représentées par le tableau I.1. Dans le cadre de cette étude, le portefeuille pourra évoluer en fonction des besoins en *run-off* ou alors en *New Business* avec la mécanique décrite précédemment dans la partie I.4.2. En *New Business*, 405 MP supplémentaires sont intégrés, présentant des caractéristiques moyennes similaires aux MP initiaux. Les contrats proposés par l'assureur sont des contrats d'assurance vie en euros et en unités de compte (UC) avec des taux minimums garantis (TMG) permettant de capturer les effets liés à un portefeuille qui s'est construit sur plusieurs générations de taux.

PASSIF	Valeur Moyenne
Age	56 ans
Ancienneté Fiscale	16 ans
PM ⁹ Ouverture Euro	35 M€
PM Ouverture UC	8,75 M€
Taux De Chargement Sur Encours Euro	0,80%
Taux De Chargement Sur Encours UC	1,05%
PB Contractuelle	95%
TMG Net	0,77%

TABLE I.1 – Valeurs moyennes au passif des MP

En ce qui concerne les proportions des contrats du portefeuille, les contrats UC représentent 20 % des provisions mathématiques tandis que les contrats euros représentent quant à eux 80 %. La provision mathématique moyenne pour les fonds Euros par MP est de 35 millions d'euros, alors que pour les UC, elle s'établit à 8,75 millions d'euros. Enfin, la table I.2 offre une vision globale des différentes variables du passif.

9. PM signifie Provision Mathématique, qui représente la valeur actuelle des engagements futurs de l'assureur envers les assurés.

	Montant
PM initiale	43,7 M€
RC¹⁰ initiale	525 000 €
PPB initiale	2,1 M€
Passif Total	46,37 M€

TABLE I.2 – Situation initiale du passif

I.6.2 L'actif

Les actifs modélisés dans le cadre de cette étude sont les suivants : les actions, les obligations, l'immobilier et le cash. Les obligations sont segmentées selon des *model points* au nombre de 60. L'ensemble des caractéristiques des *model points* obligataires sont les suivantes :

- Le type d'obligation : entreprise ou état ;
- La maturité résiduelle ;
- La valeur nominale ;
- La valeur nette comptable (VNC) ;
- La valeur de marché (VM).

Pour les autres classes d'actifs, ils ont été modélisés en un seul *model point*. Les distributions globales d'actifs sont présentées dans la table I.3.

Type d'actifs	VNC	VM
Actions	6 208 125 €	8 443 050 €
Immobilier	2 445 625 €	3 252 681 €
Obligations	28 218 750 €	26 269 775 €
État	12 792 500 €	11 769 100 €
Entreprise	15 426 250 €	14 500 675 €
Cash	752 500 €	752 500 €

TABLE I.3 – Situation initiale de l'actif

A l'instant initial, les différentes classes d'actifs affichent des taux variés de plus ou moins-values latentes (PMVL) comme l'illustre la figure I.10. Cet indicateur est crucial pour comprendre la position de l'assureur. Pour rappel, la formule est la suivante :

$$\text{PMVL} = \frac{\text{VM}}{\text{VNC}} - 1. \quad (\text{I.16})$$

10. RC signifie réserve de capitalisation, qui représente une réserve financière constituée à partir des bénéfices non distribués, utilisée pour couvrir les engagements futurs ou renforcer la solidité financière de l'entreprise.

On observe que le portefeuille détient d'importantes plus-values latentes en ce qui concerne les actions et l'immobilier (soit 36 % et 33 % respectivement), tandis que le portefeuille est en moins-value latente en ce qui concerne la poche obligataire (à -8 % en moyenne). Cette situation de moins-value latente est attribuable à la remontée brutale des taux d'intérêt car quand les taux montent le prix des obligations en portefeuille baisse.

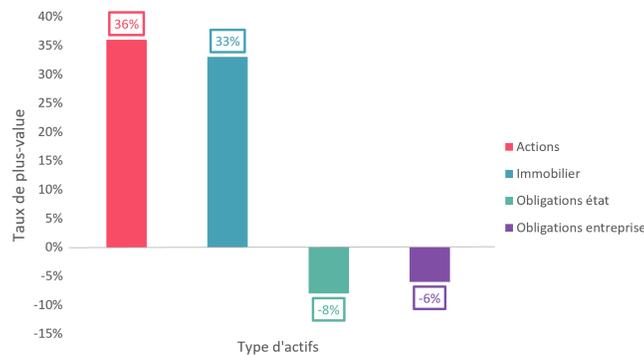


FIGURE I.10 – Taux de PMVL initiales

Pour finir, la figure I.11 décrit l'état initial du portefeuille d'actifs avec une poche obligataire d'environ 72 %, les actions pour 19 %, l'immobilier pour 7 % et enfin le cash pour 2 %. L'allocation possède un pourcentage élevé d'actions, cela est dû au fait que la stratégie était auparavant confrontée à un contexte de taux bas, ce qui permettait d'aller chercher du rendement par le biais des actions. La répartition de chaque catégorie d'actifs au sein du portefeuille est exprimée en pourcentage de l'assiette globale des actifs, calculée en valeur de marché.

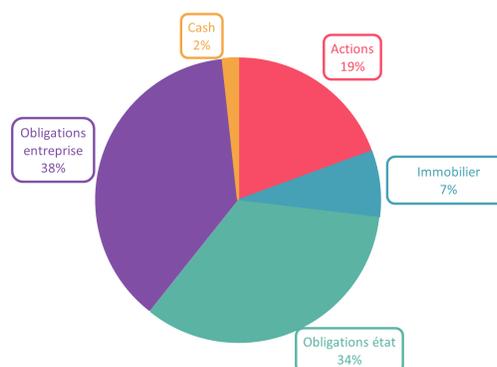


FIGURE I.11 – Répartition initiale des classes d'actifs

Après avoir décrit la composition du portefeuille, pour évaluer sa performance dans un cadre économique ou réglementaire, il est nécessaire de définir des indicateurs de pilotage pour la suite de cette étude.

I.7 Définitions des indicateurs utilisés

Les modélisations seront réalisées dans un univers risque neutre puis en risque réel. Des indicateurs annuels ou par simulations seront calculés grâce à des moyennes afin de produire des éléments de comparaison entre les différentes modélisations.

PVFP (*Present Value of Future Profits*) :

La PVFP représente la valeur actuelle des profits futurs attendus d'un portefeuille d'assurance. C'est un indicateur clé pour évaluer la rentabilité future des contrats en cours. Le résultat s'exprime à l'aide de l'équation suivante :

$$\text{Résultat} = \text{Solde}_{\text{souscription}} + \text{Solde}_{\text{financier}} - \text{Frais} \quad (\text{I.17})$$

Le solde de souscription fait référence au solde résultant des opérations d'assurance proprement dites. Il est déterminé par la différence entre les primes encaissées et les prestations.

Le solde financier, quant à lui, englobe les revenus et les coûts associés aux placements réalisés par l'assureur avec les primes encaissées.

Mathématiquement, le solde financier peut être représenté comme suit :

$$\text{Solde}_{\text{financier}} = \text{Produits}_{\text{financiers}} - \text{Intérêts techniques} - \text{PB} - \Delta\text{PPB} - \Delta\text{RC} \quad (\text{I.18})$$

avec ΔRC et ΔPPB respectivement la variation de la réserve de capitalisation et de participation aux bénéfices sur l'année.

Le calcul de la PVFP est donné par la formule suivante :

$$\text{PVFP} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^T \frac{\text{Résultat}_{i,j}}{(1 + \tau_{i,j})^j} \quad (\text{I.19})$$

où :

- N le nombre de simulations ;
- T le nombre d'années de projection ;
- $\text{Résultat}_{i,j}$ le résultat de la simulation i à l'année de projection j ;
- $\tau_{i,j}$ le facteur d'actualisation de la simulation i pour l'année j .

BEL (*Best Estimate Liability*) :

Le *Best Estimate Liability* (BEL) ou *Best Estimate* (BE) mesure l'engagement de l'assureur envers les

assurés par la somme actualisée au taux sans risque de tous les flux futurs probables induits par son portefeuille de contrats. Le BE se place dans une vision du portefeuille en *run-off*, de telle sorte que l'on considère tous les flux futurs jusqu'à l'extinction des contrats.

La provision *Best Estimate* est alors déterminée comme suit :

$$BEL = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^T \frac{CF_i(j)}{(1 + r_{i,j})^j} \quad (I.20)$$

où :

- N le nombre de simulations ;
- T le nombre d'années de projection ;
- $r_{i,j}$: le facteur d'actualisation sans risque de l'année j pour la simulation i ;
- $CF_i(j)$: les cash-flows sortants de l'année j pour la simulation i

RM (*Risk Margin*) :

La *Risk Margin* (RM), comme introduite, dans la partie I.2 de ce mémoire, est la marge de risque liée aux incertitudes de calcul du *Best Estimate*. Mathématiquement, la *Risk Margin* est calculée en utilisant la formule de coût du capital, qui est définie comme suit :

$$RM = CoC_{rate} \times \sum_{j=0}^T \frac{SCR(j)}{(1 + r_{j+1})^{j+1}} \quad (I.21)$$

avec :

- CoC_{rate} (taux de coût du capital) le taux fixé par les régulateurs, représentant le coût du capital nécessaire pour soutenir les obligations d'assurance. Sous Solvabilité II, il est établi à 6 % (ce taux a été révisé après la revoyure à 4,75 %) ;
- $SCR(j)$ est le capital de solvabilité requis à l'année j ;
- r_{j+1} représente le taux d'intérêt de base sans risque pour l'échéance de $j + 1$ an ;
- T le nombre d'années de projection

VIF (*Value In Force*) :

La *Value In Force* (VIF) correspond à la PVFP minorée de la marge pour risque introduite précédemment I.21.

$$VIF = PVFP - RM \quad (I.22)$$

La VIF permet d'évaluer la valeur d'un portefeuille de contrats de l'assureur. En d'autres termes, la VIF capture la valeur nette des bénéfices futurs attendus des contrats d'assurance, après avoir pris en compte le coût du capital pour couvrir les risques assurantiels.

TRA (Taux de Rendement de l'Actif) :

Ce taux mesure le rendement que tire l'assureur de l'ensemble des actifs qu'il détient en portefeuille. Cet indicateur permettra notamment dans un univers de projection risque réel de pouvoir évaluer les performances des stratégies d'allocation d'actifs. Le taux de rendement des actifs (TRA) se calcule annuellement de la manière suivante :

$$\text{TRA}(j) = \frac{\text{Produits}_{\text{financiers}}(j)}{\text{VNC}_{\text{totale}}(j)} \quad (\text{I.23})$$

Les produits financiers sont constitués :

- **Des PMVL** réalisées ainsi que **les turnovers**¹¹ pour l'action et l'immobilier ;
- Le montant de **la réserve de capitalisation** et des PMVL obligataires ;
- **Des coupons, l'amortissement et les remboursements** des obligations ;
- **Les intérêts** liés à l'immobilier, les actions et le cash.

La richesse latente :

La richesse latente est un indicateur financier permettant d'étudier la santé économique du portefeuille. Plus sa valeur est élevée, plus cela indique un signe positif d'une bonne santé financière pour le portefeuille. Elle est constituée de trois indicateurs, dont la PMVL introduite lors de la présentation de l'actif du portefeuille I.6.2 ainsi que la réserve de capitalisation (RC) et la provision pour participation aux bénéfices (PPB) :

$$\text{Richesse}_{\text{latente}} = \text{PMVL} + \text{RC} + \text{PPB} \quad (\text{I.24})$$

Cet indicateur permet de mesurer la richesse accumulée par l'assureur au fur et à mesure de la projection provenant de ses investissements.

Après avoir décrit l'ensemble de la structure du modèle ALM, il convient toutefois de souligner que le modèle VBA présente certaines limites. Il rencontre des contraintes temporelles et de stabilité importantes lors de simulations plus complexes, notamment en ce qui concerne les enjeux de modélisation abordés dans ce mémoire.

11. Montant de plus-value automatique

I.8 Limites du modèle VBA dans le cadre actuel

Au fil des années et avec l'évolution rapide de la technologie et des besoins, certaines limites de la version du modèle ALM VBA sont apparues, en particulier dans le secteur de l'actuariat. De plus, en envisageant l'intégration de méthodes plus avancées telles que le *reinforcement learning*, le langage montre des limites en matière de flexibilité, notamment en raison de l'impossibilité d'intégrer certaines des bibliothèques (PyTorch, Pandas, etc...) et des *frameworks* modernes, essentiels à la mise en œuvre des algorithmes d'apprentissage par renforcement.

Face à ces contraintes inhérentes au VBA, la quête d'une alternative technologique devient non seulement souhaitable mais impérative. C'est ici qu'intervient le langage de programmation Python. Ce langage riche permet de fournir les outils nécessaires à l'implémentation des modèles de *reinforcement learning*. Une conversion du modèle ALM de VBA vers Python a donc été effectuée, en conservant l'architecture du modèle initial. Une optimisation des calculs a été mise en place en s'appuyant sur des bibliothèques comme NumPy afin de traiter les données sous forme de tableaux plus rapidement.

Pour vérifier la bonne conversion du modèle, des tests de validation du modèle Python par rapport au modèle VBA seront conduits pour assurer que le modèle Python produit les mêmes résultats que le modèle VBA, mais qu'il les produit également de manière plus rapide. Ces tests seront explicités dans la partie suivante de ce mémoire.

Une fois les limites du modèle VBA dressées, il est important de tester la robustesse du nouveau modèle ALM. Pour cela, le modèle génère une synthèse d'informations constituée des indicateurs définis en I.7. La synthèse obtenue permet, en premier lieu, de valider le nouveau modèle ALM et, dans un second temps, d'étudier les performances des stratégies d'allocation d'actifs.

I.9 Validation du modèle ALM

La validation du modèle ALM est une étape cruciale. En effet, elle vise à garantir la robustesse et la fiabilité des calculs. Pour cela, une première étape de validation est la lecture du bilan Solvabilité II afin d'observer le bon balancement entre actif et passif au travers de l'écart de convergence I.7. Les paramètres de cette étude sont les suivants : projection en date du 31/12/2022 sur 40 ans et 1000 simulations dans le cadre du scénario stochastique à l'aide des scénarios générés par le GSE.

La table I.4 illustre les résultats du scénario central avec 1000 simulations au travers du bilan S2 du nouveau modèle. Les résultats permettent d'établir la robustesse du modèle, avec un écart relatif entre le passif et l'actif d'environ 0,2 %.

Actif		Passif	
Obligations	26 269 775 €	PVFP	3 449 110 €
- Etat	11 769 100 €		
- Corporate	14 500 675 €		
Actions	8 443 050 €	BEL	43 913 399 €
- Type 1	8 443 050 €		
- Type 2	-		
Immo	3 252 681 €	Ecart de convergence	105 496 €
Cash	752 500 €		
UC	8 750 000 €		
Total	47 468 006 €	Total	47 468 006 €
EC Relatif 0,2 %			

TABLE I.4 – Bilan S2 - Scénario central

La décomposition $SCR_{marché}$ est représentée à l'aide de la figure I.12 ci-dessous. Elle permet d'observer la proportion de chaque sous-modules de SCR dans le $SCR_{marché}$. La partie au-dessus du $SCR_{marché}$ correspond au gain de diversification provenant de l'agrégation des sous-modules de SCR à l'aide de la matrice de corrélation fournie par le régulateur, en formule standard. Le SCR_{action} représente plus de 50 % du module, cela nécessite donc une attention particulière lors du travail d'optimisation afin de suivre l'évolution de ce SCR.

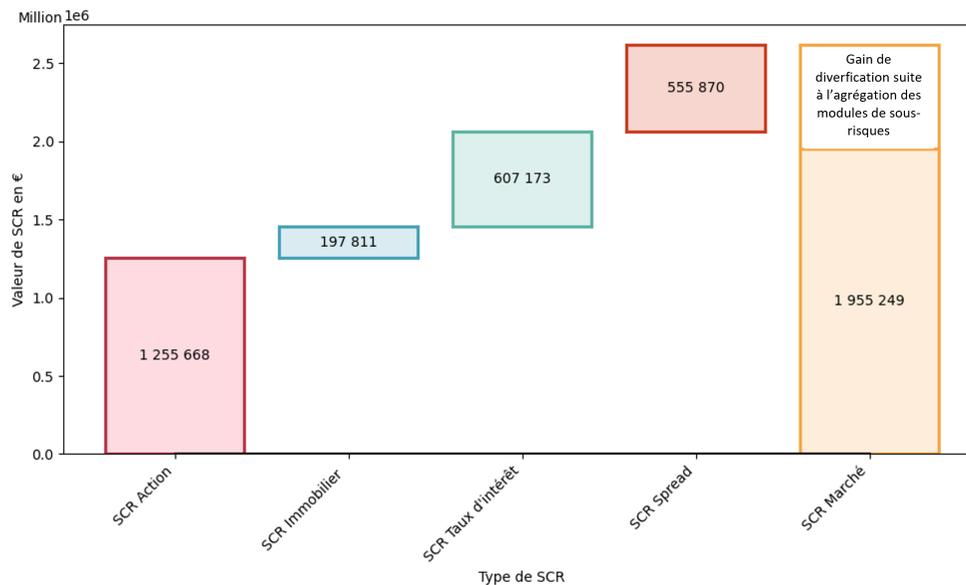


FIGURE I.12 – Décomposition du SCR marché

De plus, dans un contexte de taux élevés, le SCR_{taux} est exposé à un choc à la hausse des taux d'intérêt. Cette situation résulte d'un portefeuille constitué avec des obligations dans un environnement de taux bas, exposant ainsi le portefeuille à un risque accru en cas de hausse des taux dans le contexte du 31/12/2022.

Après avoir décrit le $SCR_{marché}$ la figure ci-dessous présente la décomposition du $SCR_{souscription}$ (cf.I.13). Ce SCR est dominé majoritairement par un SCR_{rachat} (massif). Ce choc sur les encours nécessite la liquidation d'une quantité significative de valeurs de marché en début de projection pour honorer les prestations aux assurés à des taux élevés, dans le but de limiter les rachats. Cela induit une perte d'opportunité pour l'assureur tout au long de la projection pour améliorer son résultat financier.

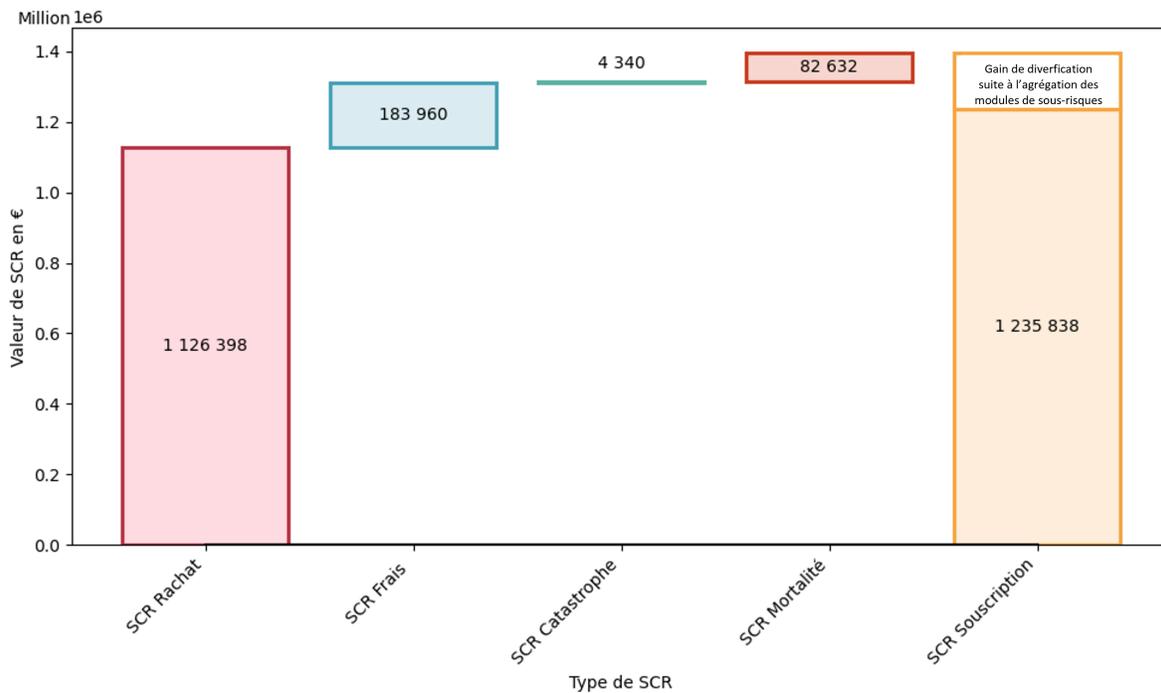


FIGURE I.13 – Décomposition du SCR souscription

Après avoir décrit de manière théorique les mécanismes de gestion actif-passif, ainsi que les indicateurs de performance et de solvabilité associés au modèle ALM, l'environnement de l'étude est dressé. L'assureur est confronté à des choix stratégiques pour satisfaire les contraintes réglementaires ainsi que pour rester attractif pour les assurés. Il a donc recours à des outils de modélisation prospective. Tout d'abord, différentes stratégies d'allocation d'actifs existantes seront étudiées et comparées. L'évaluation des avantages et des limites de chaque stratégie permettra d'établir la fondation du modèle construit. Ainsi dans la continuité de cette démarche, une approche par *reinforcement learning* sera proposée.

Chapitre II

La gestion actif-passif en assurance vie : état de l'art et lien avec le *Machine Learning*

II.1 Les différentes méthodes classiques de gestion

Les méthodes de solutions ALM sont regroupées selon deux axes : l'horizon de temps (une ou plusieurs périodes) et les facteurs de risque (déterministes ou stochastiques). Les solutions à période unique (par exemple, un an) sont des solutions à court terme et ne prennent en compte que le moment présent, tandis que les solutions à multipériodes prennent en compte les futures années de projection et permettent donc d'obtenir une vision plus long-termiste. Sur l'axe des facteurs de risque, les solutions déterministes considèrent que l'incertitude dans le modèle restera constante au fil du temps, contrairement aux solutions stochastiques. La catégorie des modèles déterministes peut être historiquement divisée en deux sous-catégories : les méthodes basées sur la notion d'« immunisation » et celles de « surplus ». Avant de présenter en détail ces méthodes, il est nécessaire d'effectuer un rappel des différentes notions découlant de ces procédés.

Taux actuariel

Le taux actuariel d'une obligation est le taux d'intérêt qui égalise la valeur actualisée de tous les flux de trésorerie futurs de l'obligation (incluant les coupons et le prix de remboursement à l'échéance du titre) au prix actuel de l'obligation sur le marché. Cela permet notamment aux compagnies d'assurance de mieux évaluer le risque et la rentabilité de leurs investissements en obligations, ainsi que de prendre des décisions éclairées sur la gestion de leur portefeuille obligataire dans un contexte d'ALM. En effet, si le taux actuariel d'une obligation est inférieur au taux d'intérêt du marché, cela signifie que l'obligation est sous-évaluée et représente une opportunité d'achat intéressante.

Le taux actuariel vérifie l'équation :

$$\text{Prix de marché } (P) = \text{Valeur actuelle (VA)} = \sum_{t=1}^n \frac{CF_t}{(1+r)^t} \quad (\text{II.1})$$

où :

- P représente le prix de marché de l'obligation,
- CF_t représente le flux de trésorerie à l'instant t ,
- r représente le taux actuariel de l'obligation,
- t représente l'instant auquel le flux de trésorerie CF_t est versé.
- n représente le nombre d'années de projection.

Duration

La duration est la durée (exprimée en années) à l'échéance à laquelle s'annihilent l'effet-coupon et l'effet-capital. La duration au sens de Macaulay [1938] est la durée moyenne pondérée jusqu'à l'échéance des flux de trésorerie d'une obligation. Le poids de chaque flux financier est déterminé en divisant la valeur actuelle du flux financier par le prix.

L'expression de la duration est la suivante :

$$\text{Duration} = \frac{\sum_{t=1}^n t \frac{CF_t}{(1+r)^t}}{\sum_{t=1}^n \frac{CF_t}{(1+r)^t}} = \frac{\sum_{t=1}^n t \frac{CF_t}{(1+r)^t}}{VA} \quad (\text{II.2})$$

Une autre façon d'interpréter cette mesure serait de la considérer comme la durée moyenne pondérée pendant laquelle un investisseur doit maintenir une position dans l'obligation jusqu'à ce que la valeur actuelle des flux de trésorerie de l'obligation soit égale au montant payé pour l'obligation.

La duration d'une obligation à coupon sera toujours inférieure ou égale à sa maturité. Les seules exceptions sont les zéro-coupons qui ont une duration égale à leur maturité.

Sensibilité

La sensibilité est un indicateur qui exprime la variation mesurable de la valeur d'un titre en réponse à une variation des taux d'intérêt. La sensibilité repose sur le concept selon lequel les taux d'intérêt et les prix des obligations évoluent dans des directions opposées. Cette formule est utilisée pour déterminer l'effet qu'une variation de 100 points de base (1 %) des taux d'intérêt aura sur le prix d'une obligation. La sensibilité est ainsi donnée par l'expression suivante :

$$\text{Sensibilité} = \frac{dVA}{VA \times dr} = -\frac{1}{VA} \times \sum_{t=1}^n t \frac{CF_t}{(1+r)^{t+1}} \quad (\text{II.3})$$

Cette formule montre que la sensibilité est inversement proportionnelle au prix de l'obligation (plus le prix est élevé, plus la sensibilité est faible) et qu'elle dépend de la dérivée partielle du prix par rapport aux taux d'intérêt.

On peut aussi définir la sensibilité en fonction de la duration :

$$\text{Sensibilité} = \frac{\text{Duration}}{1+r} \quad (\text{II.4})$$

Une interprétation intuitive de cette formule est que la sensibilité mesure la pente de la courbe de prix de l'obligation par rapport aux taux d'intérêt. Ainsi, une obligation avec une sensibilité élevée sera plus volatile et aura une plus grande variation de prix en réponse aux variations des taux d'intérêt.

II.1.1 Les méthodes d'immunisation du portefeuille

L'immunisation est, selon Redington [1952], une méthode monopériodique de gestion actif-passif notamment utilisée par les compagnies d'assurance pour minimiser leur exposition aux risques de taux d'intérêt. Elle consiste à aligner les flux de trésorerie du portefeuille d'actifs sur les engagements futurs en matière de passifs, afin de s'assurer que l'entreprise dispose des fonds nécessaires pour honorer ses engagements envers les assurés. Les compagnies d'assurance peuvent utiliser différentes techniques d'immunisation, les deux approches les plus connues sont :

- Le duration *matching*
- Le *cash-flow matching*

Le duration *matching*

La méthode de duration *matching* est une méthode déterministe de gestion de l'ALM qui consiste à évaluer les durations des actifs et des passifs. Pour accomplir une immunisation par la duration, l'investisseur doit acquérir des titres dont la duration moyenne est égale à la duration des flux du passif. L'immunisation par la duration d'un portefeuille n'est parfaite que si celle-ci est réalisée avec des instruments zéro-coupon. L'avantage de cette méthode est qu'elle est simple à comprendre et à mettre en œuvre, et elle permet de réduire le risque de taux pour l'assureur. Cependant, elle ne prend pas en compte les fluctuations des taux d'intérêt sur le long terme, ce qui peut entraîner des écarts entre les flux entrants et sortants à long terme. L'immunisation par cette méthode requiert, en pratique, un rebalancement périodique du portefeuille en réestimant à chaque fois la duration du passif, celle-ci changeant continuellement du fait du changement des taux d'intérêt et de l'écoulement du temps.

Le *cash-flow matching*

La méthode de *cash-flow matching* est une méthode d'ALM qui consiste à aligner les flux de trésorerie des actifs avec les flux de trésorerie des passifs. Le principe est de s'assurer que les flux de trésorerie des actifs correspondent aux obligations de paiement des passifs à chaque date de paiement prévue. Cette méthode est souvent utilisée pour les passifs définis tels que les rentes ou les prestations de retraite.

Mathématiquement, cela peut s'exprimer comme suit. Soit le flux net de trésorerie au moment t dénoté R_t , i.e. :

$$R_t = A_t - L_t \geq 0 \text{ pour } t = 1, 2, 3, \dots, n \quad (\text{II.5})$$

- A_t représente les actifs
- L_t représente le passif

Si les flux nets sont nuls pour chaque période, l'actif est dit parfaitement adossé au passif.

Bien que la méthode *cash-flow matching* soit très efficace pour gérer le risque de taux d'intérêt, elle présente également certaines limites. Tout d'abord, cette méthode nécessite des prévisions de flux de trésorerie très précises sur une période donnée, ce qui peut être difficile à réaliser dans un environnement économique incertain. De plus, du fait de ce suivi titre par titre et de la fluctuation du passif dans le temps lié aux comportements des assurés entre autres, le suivi peut être très coûteux en temps humain et en frais de transaction.

La méthode *cash-flow matching* est souvent considérée comme une méthode très court-termiste et ne trouve malheureusement pas d'application à long terme dans la pratique.

En raison des limites rencontrées avec le *cash-flow matching* et le *duration matching*, d'autres méthodes de gestion actif-passif ont été développées, notamment celles basées sur le concept de surplus qui sont des dérivées d'un modèle fondateur : le modèle d'optimisation de Markowitz.

II.1.2 Le modèle de Markowitz

Les techniques modernes de gestion d'actifs trouvent leur origine dans les travaux de Markowitz [1952], également connus sous le nom de théorie moderne du portefeuille. Ces travaux introduisent une méthode de gestion de portefeuille basée sur la diversification. Cette méthode vise à maximiser le rendement d'un portefeuille donné tout en minimisant son risque. Pour cela une introduction du cadre de cette théorie est nécessaire.

Supposons un marché à n actifs risqués notés $j = 1, \dots, n$. On note \tilde{R}_j le rendement aléatoire du titre j sur la période et \tilde{R} le vecteur aléatoire des rendements des n actifs. :

$$\tilde{R} = \left(\tilde{R}_1, \tilde{R}_2, \dots, \tilde{R}_n \right)' \quad (\text{II.6})$$

Le rendement μ_j espéré du titre j sur la période : $\mu_j = E(\tilde{R}_j)$ et μ le vecteur des rendements espérés des n actifs :

$$\mu = \left(\mu_1, \mu_2, \dots, \mu_n \right)' \quad (\text{II.7})$$

La covariance σ_{ij} entre le rendement de l'actif i et celui de l'actif j . Notons que $\sigma_{ii} = \sigma_i^2$ représente la variance du rendement de l'actif i .

La matrice de variance-covariance Σ de n actifs :

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix} \quad (\text{II.8})$$

La proportion w_j de richesse investit dans l'actif j et w le vecteur des poids w_j :

$$w = \left(w_1, w_2, \dots, w_n \right)' \quad \sum_{j=1}^n w_j = 1 \quad (\text{II.9})$$

Le rendement du portefeuille, noté \tilde{R}_p , est donné par la formule suivante :

$$\tilde{R}_p = (w_1, w_2, \dots, w_n) \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{pmatrix} = w' R \quad (\text{II.10})$$

De plus, il est possible de définir l'espérance de rendement du portefeuille, notée μ_p , tel que :

$$\mu_P = (w_1, w_2, \dots, w_n) \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = w' \mu \quad (\text{II.11})$$

Enfin l'écart-type du portefeuille σ_p :

$$\sigma_p^2 = \text{Var}(\tilde{R}_P) = \text{Var}\left(\sum_{j=1}^n w_j \tilde{R}_j\right) = w' \Sigma w \quad (\text{II.12})$$

Le problème posé par Markowitz consiste à trouver les proportions optimales d'investissement dans chaque titre pour atteindre ces niveaux de rendements, tout en respectant les contraintes fixées. Un portefeuille efficace, ou efficient, sera celui qui aura l'espérance de rentabilité la plus élevée parmi les portefeuilles qui respectent les contraintes et ont la même variance de rentabilité que lui.

Pour cela, le modèle de Markowitz utilise la notion de frontière efficiente, qui est représentée par la courbe dans l'espace (μ, σ^2) qui relie tous les portefeuilles possibles pour un certain niveau de rendement attendu.

La figure II.1 illustre un exemple de la frontière efficiente de Markowitz. Cette frontière efficiente est la limite supérieure de l'ensemble des points. Si un portefeuille se trouve sur cette frontière, il est considéré comme efficient.

Le modèle de Markowitz est un problème d'optimisation quadratique. Il consiste à minimiser la variance du rendement du portefeuille, sous contrainte que l'espérance du rendement du portefeuille soit égale à un rendement minimal cible qu'on notera μ_{obj} .

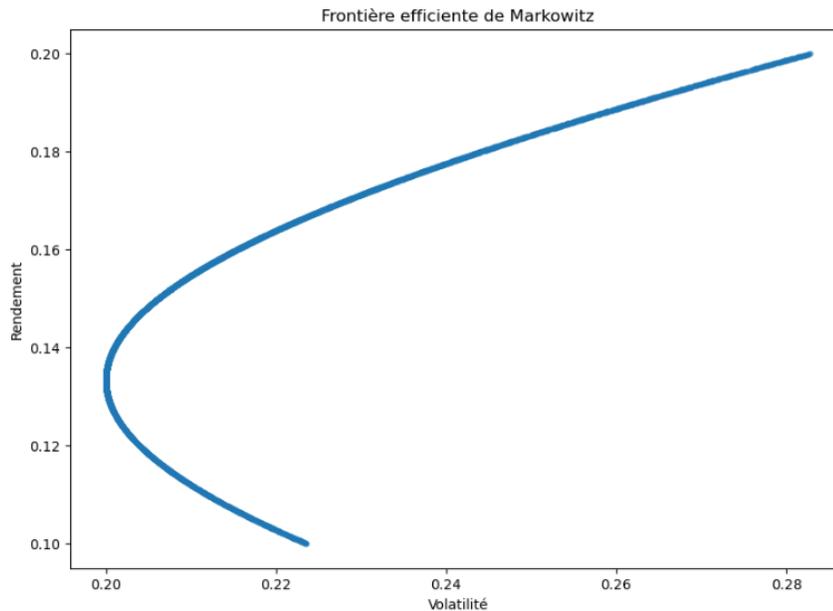


FIGURE II.1 – Frontière efficiente de Markowitz

Formellement, le problème peut être écrit comme suit :

$$\begin{cases} \min_{w \in \mathbb{R}^n} & w' \Sigma w \\ \text{s.c.} & w' \mathbf{1}_{\mathbb{R}^n} = 1 \\ & w' \mu = \mu_{obj} \end{cases} \quad (\text{II.13})$$

Bien que le modèle de Markowitz soit un outil puissant pour optimiser l'allocation des actifs, il présente plusieurs limites importantes. Premièrement, le modèle suppose que les investisseurs puissent prendre des positions courtes sans restriction, ce qui n'est pas toujours possible en pratique. De plus, le modèle n'autorise aucune transaction après l'investissement initial pendant toute la durée de la période, ce qui ne reflète pas la réalité de la gestion dynamique des portefeuilles.

De plus, le modèle repose sur l'inversion de la matrice de variance-covariance, ce qui peut poser des problèmes en pratique. L'estimation de cette matrice peut être difficile, et son inversion peut être numériquement instable si la matrice n'est pas bien conditionnée.

En outre, la stratégie ne tient pas compte des informations ou des sentiments particuliers de l'investisseur, ce qui peut être une limite importante dans la pratique. Par exemple, un investisseur peut avoir des croyances personnelles sur certains actifs qui ne sont pas prises en compte dans le modèle.

Enfin, le modèle de Markowitz repose sur l'hypothèse d'un marché efficient¹, ce qui n'est pas le cas dans la réalité. Par exemple, il ne prend pas en compte les coûts de transaction, les taxes, les restrictions

1. L'hypothèse d'efficience du marché est que les prix et les rendements des actifs intègrent de manière objective et complète toutes les informations disponibles les concernant.

réglementaires, ou l'impact des transactions de l'investisseur sur les prix du marché.

Les modèles de surplus interviennent précisément à ce stade. Ces modèles, tels que le modèle de Sharpe et Tint [1990], vont au-delà de l'optimisation du rendement des actifs et visent à optimiser le surplus.

II.1.3 Les modèles de surplus

En matière de gestion actif-passif, plusieurs modèles ont été développés pour traiter la notion du surplus, qui est généralement défini comme la différence entre la valeur de marché des actifs et la valeur actuelle du passif. Parmi les modèles les plus connus, on peut citer celui de Kim et Santomero [1988], Sharpe et Tint [1990] et Leibowitz [1992].

Le modèle de Sharpe et Tint est l'un des plus utilisés. C'est un modèle d'optimisation qui cherche à minimiser la variance de la rentabilité du surplus pour un niveau de rentabilité du surplus donné et sous un ensemble de contraintes. Dans ce modèle, le surplus à chaque instant t est donné par la formule :

$$S_t = A_t - m * L_t \quad (\text{II.14})$$

où :

- S_t est le surplus à l'instant t ;
- A_t est la valeur des actifs à l'instant t ;
- L_t est la valeur des passifs à l'instant t ;
- m correspond au poids que l'assureur accorde au passif. Dans la suite de ce mémoire, afin de simplifier les notations et de souligner l'importance du passif en assurance-vie, on suppose que $m = 1$.

Le modèle est une approche monopériodique visant à optimiser le surplus entre deux points dans le temps, en posant $t=0$ à l'origine et $t=1$ à un moment futur. Le surplus en $t=1$ se calcule de la manière suivante :

$$\widetilde{S}_1 = A_0 [1 + \widetilde{R}_p] - L_0 [1 + \widetilde{R}_L] \quad (\text{II.15})$$

Où \widetilde{R}_p est le rendement du portefeuille de l'actif (entre $t=0$ et $t=1$) et \widetilde{R}_L est le taux d'accroissement du passif (entre $t=0$ et $t=1$).

La rentabilité du surplus (R_s) est alors définie comme la variation du surplus rapportée à la valeur initiale des actifs :

$$R_s = \frac{\widetilde{S}_1 - S_0}{A_0} \quad (\text{II.16})$$

Le programme d'optimisation du modèle de Sharpe et Tint est le suivant :

$$\begin{cases} \min_{w \in \mathbb{R}^n} \text{Var}[R_s] \\ \text{s.c. : } \begin{cases} E[R_s] = r_s \\ w' \mathbf{1}_{\mathbb{R}^n} = 1 \end{cases} \end{cases} \quad (\text{II.17})$$

L'ensemble des portefeuilles répondant à ces conditions forme la frontière efficiente dans le cadre du modèle. Dans leur modèle, Sharpe et Tint prennent aussi en considération une contrainte de déficit qui correspond à la probabilité que la rentabilité du surplus soit inférieure à un certain seuil. Cette probabilité ne doit pas dépasser une certaine valeur définie par l'utilisateur.

La contrainte de déficit est représentée par une droite dans le plan de la rentabilité espérée contre la volatilité (ou le risque). Cette droite est déterminée par l'équation suivante :

$$E(R_s) = \sigma_t \cdot N^{-1}(1 - p) + u \quad (\text{II.18})$$

où :

- $E(R_s)$ est la rentabilité espérée du surplus ;
- σ_t est la volatilité (l'écart-type) du surplus ;
- $N^{-1}(1 - p)$ est le quantile de la loi normale centrée réduite correspondant à une probabilité $1 - p$;
- u est le seuil de rendement du surplus.

La droite de la contrainte de déficit montre donc la combinaison acceptable de rentabilité espérée et de volatilité pour un niveau de risque de déficit donné. Lorsque cette droite est superposée à la frontière efficiente, l'intersection de ces deux courbes comme le montre la figure II.2 donne le portefeuille efficient qui maximise la rentabilité pour un niveau de risque donné, tout en respectant la contrainte de déficit.

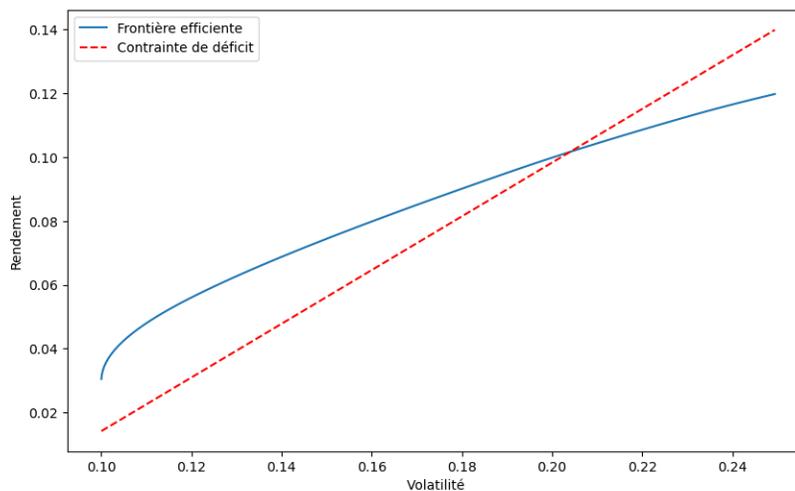


FIGURE II.2 – Frontière efficiente et contrainte de déficit de Sharpe & Tint

Malgré la sophistication et l'efficacité du modèle de Sharpe et Tint, celui-ci présente certaines limites. Comme le modèle de Markowitz, le modèle de Sharpe et Tint est monopériodique, c'est-à-dire qu'il effectue des projections que pour une seule période de temps à venir.

Face à ces limitations des modèles déterministes, les modèles stochastiques d'ALM ont été développés pouvant être utilisés pour une multitude de périodes. Ces modèles tentent d'apporter des solutions plus flexibles et plus complètes pour la gestion actif-passif.

II.1.4 Les modèles stochastiques

La simulation stochastique est une technique numérique qui permet de modéliser l'incertitude et la variabilité dans les systèmes complexes. Les simulations de Monte-Carlo sont des algorithmes qui servent à générer un grand nombre de scénarios de rendement des actifs en utilisant des distributions de probabilités (ces distributions sont généralement établies sur l'historique des rendements de l'actif). Chaque scénario est ensuite utilisé pour calculer le rendement du portefeuille en effectuant la moyenne des rendements des scénarios, et les résultats sont utilisés pour estimer la distribution des rendements futurs du portefeuille.

Les stratégies stochastiques sont des stratégies financières qui prennent en compte la nature aléatoire des marchés financiers en s'appuyant sur les méthodes de simulations stochastiques. Dans l'ouvrage de Perold et Sharpe [1988], on distingue deux types de familles de modèles : les modèles à poids statiques et les modèles à poids dynamiques. L'une des stratégies stochastiques à poids statiques les plus couramment utilisées est la stratégie de "*Fixed-Mix*".

La stratégie *Fixed-Mix*, également connue sous le nom de stratégie d'allocation d'actifs constante, est une méthode de gestion de portefeuille qui maintient une allocation d'actifs fixe et constante dans le temps, quelles que soient les conditions de marché.

Dans une stratégie de portefeuille *Fixed-Mix*, le pourcentage de chaque actif dans le portefeuille est fixé à l'avance et reste constant au fil du temps. Toutefois, en raison des fluctuations du marché, les valeurs de marché réelles des actifs dans le portefeuille changent, ce qui peut entraîner une divergence par rapport à l'allocation cible. Par conséquent, le portefeuille doit être rééquilibré périodiquement pour maintenir l'allocation cible.

Le rééquilibrage consiste à acheter ou vendre des actifs afin de maintenir l'allocation d'actifs à des niveaux prédéterminés.

Supposons un portefeuille *Fixed-Mix* qui est composé de 60 % d'actions et de 40 % d'obligations.

Dans cette stratégie, si les actions tendent à bien performer, augmentant ainsi leur part relative au sein du portefeuille, cela impliquerait de vendre une partie des actions pour revenir à l'allocation initiale de 60 %. Inversement, si les actions sous-performent et leur valeur relative diminue, il est alors nécessaire d'acheter plus d'actions pour maintenir l'allocation à 60 %.

Les implications de ces ajustements sont les suivantes :

1. **Achat bas, vente haut** : En rééquilibrant le portefeuille, cela se traduit par acheter des actifs qui ont sous-performé (et dont le prix est donc plus bas) et vendre des actifs qui ont surperformé (et dont le prix est donc plus élevé). C'est en accord avec la maxime classique de l'investissement "acheter bas, vendre haut" ;
2. **Contrôle du risque** : En maintenant une allocation fixe d'actifs, un niveau de risque constant pour le portefeuille est maintenu. Cela peut aider à éviter des pertes excessives si une classe d'actifs se déprécie fortement ;
3. **Coûts de transaction** : Le rééquilibrage du portefeuille peut entraîner des coûts de transaction, qui peuvent réduire le rendement net du portefeuille.

La stratégie *Fixed-Mix* a plusieurs avantages et inconvénients. L'un des principaux avantages est sa simplicité. Une fois que l'allocation d'actifs cible est déterminée, il ne reste plus qu'à la rééquilibrer périodiquement. De plus, le rééquilibrage régulier peut potentiellement améliorer les rendements du portefeuille, car il oblige à vendre des actifs qui ont bien performé et à acheter des actifs qui ont sous-performé, ce qui est en ligne avec la stratégie d'investissement classique "acheter bas, vendre haut".

Cependant, bien que la *Fixed-Mix* assure une certaine stabilité des rendements financiers tout au long de la période de projection, elle présente certaines limites. Notamment, cette stratégie ne tient pas compte des variations potentielles des conditions de marché et des opportunités d'investissement. En d'autres termes, elle ne permet pas de réagir de manière proactive aux changements de marché.

Dans le cadre de cette stratégie, les réinvestissements en obligations se font sur des titres dont les échéances varient entre 1 et 20 ans, tout au long de la période de projection. Cependant, pour un assureur, il n'est pas nécessairement judicieux d'investir dans des obligations à long terme en fin de période de projection. En dépit de cela, la stratégie *Fixed-Mix* peut mener l'assureur à investir dans des obligations à longue échéance durant les dernières années de la projection.

Cela implique que l'assureur devra revendre ces obligations acquises précédemment lors de la liquidation du portefeuille quelques années plus tard. Ce processus augmente le risque de réalisation de moins-values, car la revente des obligations pourrait se faire à un moment où leur valeur de marché est inférieure au prix d'achat. Ainsi, cette stratégie peut potentiellement engendrer des risques supplémentaires pour l'assureur.

Compte tenu de ces limitations inhérentes à la stratégie *Fixed-Mix*, il devient nécessaire d'explorer d'autres approches de gestion de portefeuille qui permettent une plus grande flexibilité et adaptabilité aux conditions changeantes du marché. C'est là qu'interviennent les modèles dynamiques stochastiques.

II.1.5 Les modèles dynamiques stochastiques

La principale problématique liée aux modèles à poids constant comme évoqué précédemment est leur manque d'adaptabilité et leur incapacité à capter les évolutions haussières du marché. Ainsi, sont nées

les stratégies d'allocation dynamique dont une stratégie qui est l'assurance du portefeuille ou encore *Constant Proportion Portfolio Insurance* (CPPI) qui permet de garantir un montant minimal tout en profitant des effets leviers du marché.

La méthode CPPI est basée donc sur une allocation dynamique entre un actif risqué (généralement des actions) et un actif sans risque (obligations, fonds monétaires...). L'allocation est ajustée en fonction de la valeur du portefeuille afin de maintenir un montant minimal appelé "plancher" de sécurité afin de protéger l'investisseur face aux baisses importantes du marché. En fonction de l'évolution des actifs risqués sur le marché, des contraintes mécaniques d'achat et de vente sont appliquées en fonction de paramètres prédéfinis au préalable.

On introduit les différentes terminologies associées à cette méthode :

- **Le coefficient multiplicateur** ou **multiple** dépend de la nature de la classe d'actifs à laquelle on recherche une exposition, on définit un nombre fixe m qui détermine la proportion de l'actif risqué dans le portefeuille. Un multiple plus élevé signifie une plus grande exposition au risque.
- **Le plancher** représente la fraction de la valeur liquidative du portefeuille que le gérant ne peut pas se permettre de perdre. Il est égal à la valeur actualisée du niveau de valeur liquidative garantie à l'échéance :

$$Plancher = Valeur\ Liquidative * (1 + r)^{-T} \quad (II.19)$$

avec r la taux d'actualisation et T le temps d'horizon.

- **Le coussin**, quant à lui, est la part d'actifs qu'il est possible d'investir sur des actifs risqués sans que soit remise en question la garantie :

$$Coussin = Valeur\ Liquidative - Plancher \quad (II.20)$$

- **L'exposition** est la proportion du portefeuille qui est investie en actif risqué :

$$Exposition = m * Coussin \quad (II.21)$$

À chaque variation de l'actif risqué, le multiple effectif (notons m_t), s'écarte de sa valeur cible m . L'acte de gestion consistera donc à rebalancer l'actif risqué et l'actif non risqué de façon à obtenir l'égalité entre le multiple effectif et le multiple cible. Par exemple, si le multiple effectif est inférieur au multiple cible, la stratégie CPPI vise à acheter l'actif risqué et vendre l'actif non risqué, et inversement.

Ainsi, la stratégie CPPI permet de participer à la hausse du marché tout en limitant les pertes en cas de baisse du marché. Cependant, elle nécessite un rééquilibrage régulier du portefeuille, ce qui peut entraîner des coûts de transaction.

Ces méthodes abordées pour la gestion de l'ALM ont longtemps été efficaces et largement utilisées dans

l'industrie, mais il est important de noter qu'elles possèdent certains défauts. À mesure que l'environnement financier et réglementaire se complexifie, ces approches traditionnelles montrent leurs limites. Notamment avec la réforme Solvabilité 2, l'utilisation de modèles déterministes présente plusieurs limites dans ce contexte car ils manquent de précision et supposent les conditions de marchés constantes pendant la période d'évaluation ce qui peut s'avérer être une hypothèse lourde de conséquences et ainsi conduisant à une sous-estimation ou une sur-estimation des besoins en capitaux exigés sous Solvabilité 2. Finalement, avec ces différentes stratégies, il paraît pertinent de se tourner vers les stratégies stochastiques permettant une allocation dynamique. Mais celles-ci sont encore peu répandues à cause de leurs complexités techniques et en temps de calcul.

Par ailleurs, les améliorations technologiques offrent de nouvelles perspectives pour surmonter ces obstacles, comme l'illustre le tableau II.1. C'est dans ce cadre que nous allons maintenant nous pencher sur une alternative aux méthodes classiques d'ALM et explorer des approches plus modernes, notamment celles basées sur le *machine learning* et le *deep learning*. Ces méthodes peuvent offrir des solutions plus adaptées à l'environnement actuel. Cette approche sera comparée avec la stratégie de modèle ALM existante, le *Fixed-Mix*, au travers de différents indicateurs qui seront détaillés dans la suite de ce mémoire.

Stratégie	Nature	Avantages	Inconvénients
Markowitz	Déterministe, Statique	Optimise le rapport risque/rendement	Hypothèse d'un marché parfait
Modèle de surplus	Déterministe, Dynamique	Réduit le risque de sous-capitalisation	Nécessite une connaissance précise des obligations futures
Fixed-Mix	Stochastique, Statique	Facile à mettre en œuvre	Ne s'adapte pas aux conditions de marché changeantes
CPPI	Stochastique, Dynamique	Protection contre les baisses de marché importantes	Peut entraîner des coûts de transaction élevés

TABLE II.1 – Avantages et inconvénients des différentes stratégies d'allocation de portefeuille

II.2 Le *reinforcement learning* pour la résolution d'un problème ALM

Un problème d'ALM peut être résolu grâce à la stratégie d'allocation d'actifs en utilisant une approche basée sur des techniques de *machine learning*. Dans le *machine learning*, il existe trois grands groupes d'algorithmes :

- **Les algorithmes d'apprentissage supervisé** qui cherchent à trouver des motifs, *des patterns* sur des jeux de données étiquetées.
- **L'apprentissage non supervisé** qui fait de même mais pour des données non étiquetées.
- **L'apprentissage par renforcement** qui consiste à trouver une *policy* qui maximise une certaine métrique, i.e, définir quelles actions devraient être prises dans une situation particulière afin de maximiser la probabilité que quelque chose désirée se produise.

L'application de la notion du *reinforcement learning* (RL) est le sujet central de ce mémoire. Le but est d'entraîner un algorithme qui permet, à chaque pas de temps, avec un certain montant d'actif, de passif et un jeu disponible d'investissements, de créer une allocation qui maximise la probabilité que toutes les futures dettes soient payées en respectant les contraintes réglementaires. Cette structure ci-dessus peut être modélisée comme un processus de décision Markovien (PDM). Selon Rodrigues Fontoura [2020], le RL comprend des algorithmes qui sont conçus spécifiquement afin de pouvoir répondre à des problématiques ALM. Les concepts de base du cadre d'apprentissage par renforcement seront introduits. Les processus de décision de Markov, nécessaires à la modélisation de ce problème, seront aussi décrits par la suite.

II.2.1 Concepts clés du *reinforcement learning*

Le RL consiste à apprendre à associer des situations à des actions afin de maximiser la récompense à long terme. L'apprenant ne reçoit pas d'instructions sur les actions à entreprendre, mais doit explorer l'environnement et trouver les actions les plus gratifiantes par essais et erreurs.

L'agent et l'environnement sont les composants essentiels du RL. L'environnement est le cadre dans lequel l'agent opère et interagit. L'agent décide d'une action à entreprendre après avoir pris connaissance d'une observation (potentiellement incomplète) de l'état du monde à chaque étape de l'interaction. L'environnement peut changer de lui-même ou à la suite des actions de l'agent. L'agent reçoit également une récompense de la part de son environnement, qui est un indicateur numérique de la qualité ou de la gravité de l'état actuel du monde. L'objectif de l'agent est de maximiser sa récompense totale. Les méthodes d'apprentissage par renforcement permettent à l'agent d'adopter des comportements qui l'aideront à atteindre son objectif.

Pour parler plus précisément du RL, il est nécessaire d'introduire les concepts suivants :

État

L'état décrit l'ensemble des informations actuellement accessibles pour l'agent, reflétant sa situation présente au sein de l'environnement. Il est généralement noté par $s \in S$, où S est l'ensemble de tous les états possibles. Lorsque l'agent est capable d'observer l'état complet de l'environnement, on dit que l'environnement est entièrement observé. Lorsque l'agent ne peut observer qu'une partie de l'environnement, on dit que l'environnement est partiellement observé.

Action

Une action est une décision que l'agent exécute pour interagir avec l'environnement. Elle est généralement notée par $a \in A(s)$, où $A(s)$ est l'ensemble de toutes les actions possibles dans l'état s .

Politique (*policy*)

Une politique est une règle donnée pour un agent afin qu'il puisse décider quelles actions il doit prendre. Cela peut être assimilé au cerveau de l'agent. Elle peut être déterministe et dans ce cas elle sera notée μ :

$$a_t = \mu(s_t)$$

ou elle peut être stochastique, auquel cas elle est généralement désignée par π :

$$a_t \sim \pi(a_t | s_t)$$

En stochastique, cela désigne la probabilité que l'agent joue a dans l'état s à l'instant t , i.e. :

$$P(A_t = a | S_t = s)$$

Récompense (*reward*)

Cette information désigne la fonction de récompense R que l'agent reçoit après avoir pris une action. Elle indique la qualité de l'action et est généralement notée par $r_{t+1} = R(s_t, a_t, s_{t+1})$ après avoir pris l'action a_t dans l'état s_t afin d'arriver à l'état s_{t+1} .

Les algorithmes de RL fonctionnent à travers des interactions répétées comme le montre la figure II.3 entre l'agent et l'environnement. L'agent observe un état s_t , et en fonction de sa politique, il va opérer une action a_t . À partir de cette paire état-action, l'environnement transite vers un nouvel état s_{t+1} et retourne un signal de récompense r_{t+1} . Ce processus est répété tant qu'une condition terminale n'est

pas atteinte. La séquence $s_0, a_0, r_1, a_1, r_2, ..$ générée à travers ces interactions est appelée trajectoire τ .

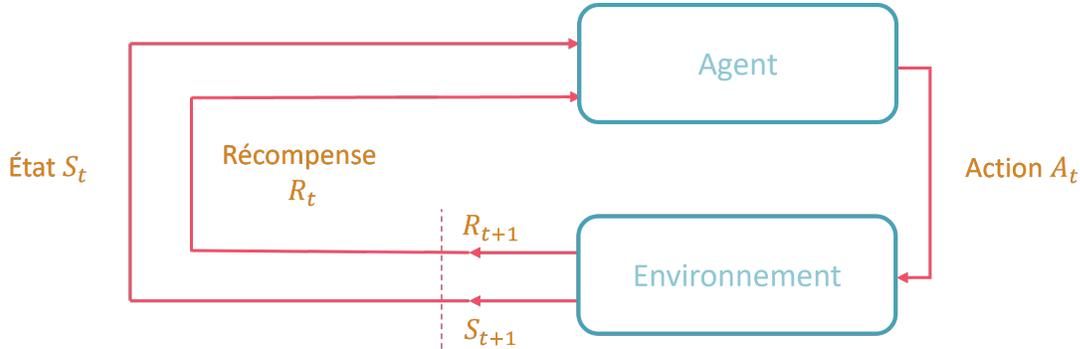


FIGURE II.3 – Interaction Agent/Environnement dans le cadre du *reinforcement learning* selon Sutton et Barto [2017]

Après avoir introduit les différents concepts liés au RL, cette structure comme introduite précédemment peut être modélisée comme un processus de décision Markovien.

II.2.2 Formalisation du processus de décision Markovien

Un processus de décision Markovien est défini par "une formalisation d'une séquence de prise de décisions, où les actions n'influencent pas immédiatement sur les récompenses, mais aussi les situations ultérieures, ou états" selon Sutton et Barto [2017].

La formalisation d'un processus de décision Markovien repose sur cinq éléments, formant un 5-tuple, $\langle S, A, R, P, \rho_0 \rangle$, où :

- S est l'ensemble des états,
- A est l'ensemble des actions valides,
- $P : S \times A \rightarrow \mathcal{P}(S)$ est la probabilité de transition, avec $P(s' | s, a)$ la probabilité de transiter vers un état s' si on commence dans un état s et avec une action a ,
- $R : S \times A \times S \rightarrow \mathbb{R}$ est la fonction de récompense, avec $r = R(s, a, s')$,
- s_0 est la distribution de l'état de départ avec une densité $\rho_0(s_0)$,

La politique $\pi_\theta : S \rightarrow \mathcal{A}$ est définie, elle sélectionne les actions à prendre dans un état donné. Elle peut être stochastique (définie par $\pi_\theta(a_t | s_t)$) ou déterministe (définie par $a_t = \mu_\theta(s_t)$). Dans les deux cas, $\theta \in \mathbb{R}^n$ est le vecteur de paramètres de cette fonction.

Le cadre formel de l'apprentissage par renforcement étant défini, la prochaine étape de la modélisation se concentre sur la manière dont l'agent peut optimiser ses actions au sein de l'environnement. L'optimisation des décisions de l'agent est un aspect crucial du RL. En effet, l'objectif de tout agent est de maximiser la récompense obtenue de son environnement. Il est nécessaire alors d'établir des métriques afin d'évaluer la valeur d'un état ou d'une action.

II.2.3 Fonctions importantes

L'objectif de l'agent est de maximiser la notion de récompense cumulative sur une trajectoire. Le rendement d'une trajectoire τ notée $G(\tau)$ ², ou encore récompense cumulée actualisée, est définie comme la somme des récompenses actualisées. Le facteur d'actualisation noté $\gamma \in [0, 1]$ est utilisé pour éviter les problèmes liés aux trajectoires infinies et pour accentuer le fait que la récompense actuelle est plus importante que les récompenses futures. Le rendement d'une trajectoire est donné par :

$$G(\tau) = \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \dots + \gamma^T r_T = \sum_{t=1}^T \gamma^t r_t \quad (\text{II.22})$$

L'objectif principal d'un algorithme de RL est d'identifier une stratégie qui maximise l'espérance en matière de récompense $\mathbb{E}[G(\tau)]$.

Cet objectif de maximisation de rendement ne peut être atteint que si l'agent peut évaluer la valeur d'une action ou d'un état, pour cela l'agent a recours aux fonctions de valeurs.

La fonction de valeur d'une politique π et d'un état s_t se note $V^\pi(s_t)$, elle représente la récompense cumulative actualisée attendue d'un état s_t avec une politique π suivie par l'agent. Cela permet de déterminer le rendement dans la situation où l'agent est dans un état s_t et suit la politique π jusqu'à la fin. On note la fonction de valeur :

$$V^\pi(s_t) = \mathbb{E}_\pi(G_t | s_t) \quad (\text{II.23})$$

Une fonction similaire est la fonction d'action-valeur $Q^\pi(s_t, a_t)$ d'un état s_t , d'une action a_t et d'une politique π . La différence par rapport à la fonction de valeur vient du fait que dans son premier pas l'agent prend une action arbitraire et ensuite suit la politique π . Cela permet d'évaluer le rendement dans la situation où l'agent prend une action maintenant dans un état s_t et suit la politique π à partir d'un état s_{t+1} jusqu'à la fin. On note la fonction d'action-valeur :

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi(G_t | s_t, a_t) \quad (\text{II.24})$$

Résoudre un problème par renforcement réside dans le fait de trouver une politique π qui est supérieure ou égale aux autres politiques, c'est la politique optimale. Elle permet de maximiser la probabilité d'atteindre un objectif de récompense sur le long terme. Une politique π est meilleure ou égale à une autre politique π' si son rendement attendu est supérieur ou égal à celui des autres politiques pour tous les états :

$$\pi \geq \pi' \Leftrightarrow V_\pi(s) \geq V_{\pi'}(s)$$

2. Bien que les concepts soient similaires, il ne s'agit pas du taux utilisé pour actualiser un passif à sa valeur actuelle.

La fonction optimale de valeur d'état est définie par $V_*(s)$:

$$V_*(s) = \max_{\pi} V_{\pi}(s)$$

La fonction optimale action-valeur est définie par $Q_*(s, a)$:

$$Q_*(s, a) = \max_{\pi} Q_{\pi}(s, a)$$

Après avoir introduit les fonctions de valeur, nécessaires pour évaluer la performance d'un agent dans un environnement donné, il est essentiel de comprendre comment ces fonctions sont calculées et mises à jour. La fonction de valeur optimale est liée récursivement aux équations de Bellman.

II.2.4 Équations de Bellman

Les équations permettent de mettre à jour les différentes fonctions de valeurs. En effet, les équations de Bellman sont des équations récursives qui, dans le cas du RL permettent d'exprimer la valeur d'un état (ou d'une paire état-action) en termes de valeurs d'autres états (ou paires état-action). Cela établit une relation entre la valeur d'un état et les états futurs qui lui succèdent.

Pour la fonction de valeur d'état, l'équation de Bellman est donnée par :

$$V^{\pi}(s) = \sum_{a \in A} \pi(a|s) \sum_{s', r} P(s', r|s, a) [r + \gamma V^{\pi}(s')] \quad (\text{II.25})$$

Pour la fonction de valeur d'action, l'équation de Bellman est donnée par :

$$Q^{\pi}(s, a) = \sum_{s', r} P(s', r|s, a) [r + \gamma \sum_{a' \in A} \pi(a'|s') Q^{\pi}(s', a')] \quad (\text{II.26})$$

Dans le cas où l'environnement est totalement connu, il s'agit d'un problème de planification, qui peut être résolu par programmation dynamique. Malheureusement, dans la plupart des scénarios, $P(s', r|s, a)$ ou r ne sont pas connues à l'avance. Il est donc impossible de résoudre les PDM en appliquant directement les équations de Bellman, mais cette méthode fixe les jalons des bases théoriques de nombreux algorithmes RL.

II.2.5 Méthodes de résolution par renforcement : *Value-based*

Un point crucial lors de la résolution d'un problème de RL est de déterminer quel type de méthode de résolution de problème de *reinforcement learning* serait le plus approprié à employer. Il existe deux approches principales pour résoudre ces problèmes : les méthodes dites *Value-based* et Optimisation de la politique.

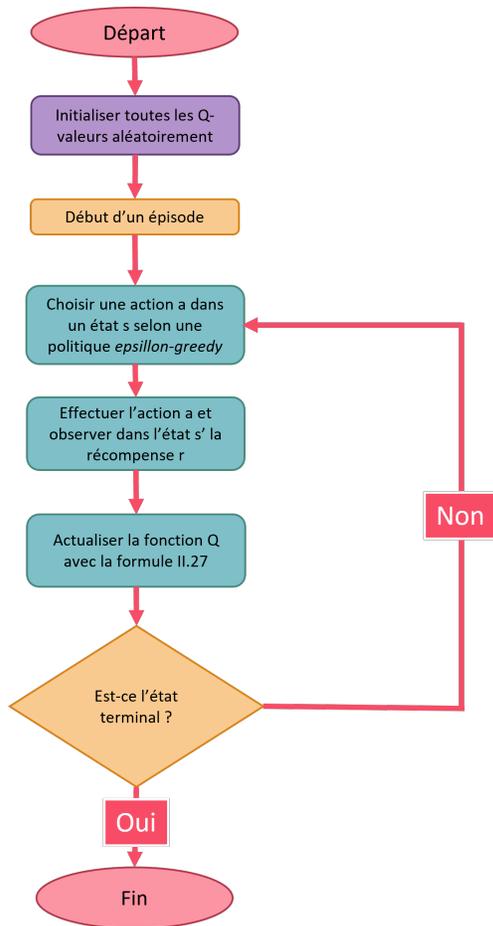
L'algorithme *Q-Learning*

Les méthodes *Value-based* ont pour objectif de trouver une stratégie optimale en s'appuyant sur la fonction d'action-valeur afin d'estimer la récompense. Le *Q-learning* est un algorithme ayant pour objectif de trouver la politique optimale en mettant à jour la fonction action-valeur (Q) à chaque pas en utilisant les équations de Bellman jusqu'à ce que cette fonction Q converge vers une fonction optimale Q_* .

La fonction est mise à jour à l'aide de la formule suivante :

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (\text{II.27})$$

Ici, α est le taux d'apprentissage, γ est le facteur d'escompte et $\max_a Q(s_{t+1}, a)$ est la valeur Q maximale pour le nouvel état sur toutes les actions possibles.



La figure II.4 illustre l'algorithme *Q-Learning*, un pseudo-code est disponible en Annexe 1.

Au travers du choix de l'action a à adopter l'agent utilise une politique dite "*epsilon-greedy*" qui instaure le principe de l'exploration et l'exploitation fondamentale dans le RL. Avec une probabilité ϵ définie en paramètre, l'agent choisit une action aléatoire (exploration), et avec une probabilité $1 - \epsilon$, l'agent choisit l'action maximisant la valeur Q (exploitation). Cette méthode permet à l'agent d'explorer de nouvelles actions afin d'éviter une solution sous-optimale tout en exploitant ses connaissances actuelles. Les actions prises dans la phase d'exploitation sont données par :

$$a(s) = \arg \max_a Q(s, a) \quad (\text{II.28})$$

FIGURE II.4 – L'algorithme *Q-Learning*

L'algorithme du *Q-Learning* est une méthode qui fonctionne bien pour trouver la politique optimale pour les problèmes avec un espace d'états et d'actions relativement petit. Mais il devient inutilisable pour les problèmes avec un grand nombre d'états ou d'actions. Pour cela, les réseaux de neurones sont capables d'opérer dans ces situations.

II.2.6 Les réseaux de neurones

Un neurone artificiel

Un neurone, dans le cadre des réseaux de neurones artificiels comme représentés par la figure II.5, est une unité de calcul qui prend en entrée un certain nombre de valeurs, effectue un calcul sur ces valeurs, puis produit une seule valeur de sortie.

La formulation mathématique d'un neurone est généralement la suivante, soit $x = (x_1, x_2, \dots, x_n)$ le vecteur d'entrées du neurone, $w = (w_1, w_2, \dots, w_n)$ le vecteur des poids associés à ces entrées, b le biais du neurone, ainsi que f une fonction dite d'activation du neurone. La sortie y du neurone est alors donnée par :

$$y = f\left(\sum_{i=1}^n w_i x_i + b\right)$$

Le fonctionnement d'un neurone peut être schématisé de la manière suivante :

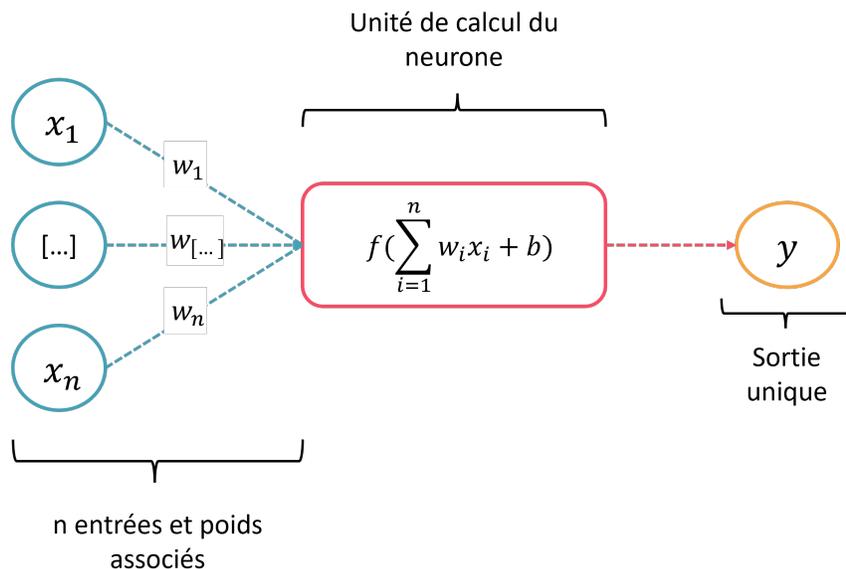


FIGURE II.5 – Représentation d'un neurone du point de vue théorique selon (DISERBEAU, 2019)

On calcule ainsi la sortie d'un neurone en effectuant la somme des entrées pondérées des poids associés,

en ajoutant un biais et en appliquant une fonction d'activation f .

La fonction d'activation permet de modéliser des relations complexes entre les entrées et les sorties du neurone. En effet, sans la fonction d'activation, le réseau de neurones ne serait qu'un simple modèle de régression linéaire.

Les couches de neurones

L'intérêt d'un neurone artificiel réside dans le fait de pouvoir agréger plusieurs itérations de neurones afin de former des couches de neurones comme illustré par la figure II.6 .

La structure typique d'un réseau de neurones entièrement connecté se compose de trois types de couches : une couche d'entrée, une ou plusieurs couches cachées, et une couche de sortie.

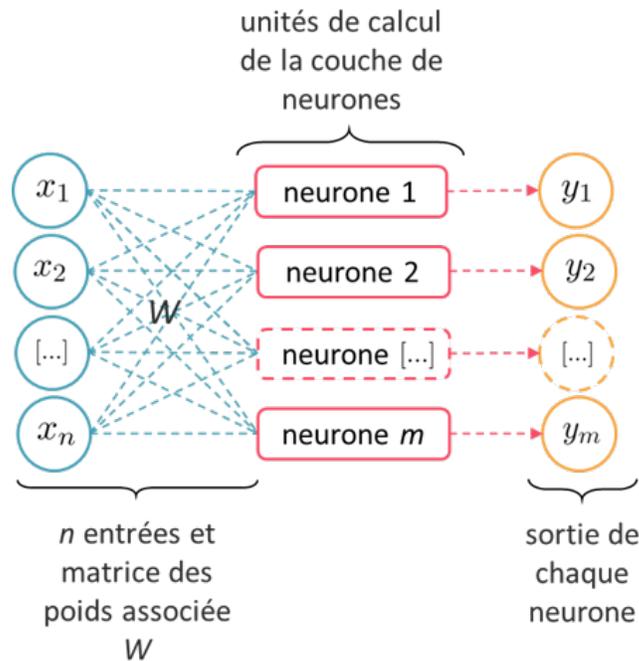


FIGURE II.6 – Représentation d'une couche de neurones

Le principe selon lequel l'agrégation de m neurones en une seule couche permet de modéliser des problèmes plus complexes peut également être appliqué aux couches de neurones elles-mêmes. En effet, il est possible de fusionner p de ces couches entre elles ce qui constitue un réseau de neurones comme ainsi présenté par la figure II.7. Ce réseau est dit *fully-connected* si les couches sont entièrement connectées c'est-à-dire que chaque neurone d'une couche est relié à chaque neurone de l'autre couche. Le calcul de la sortie d'un neurone dans son réseau est issu du même procédé que pour le neurone seul.

On généralise le calcul de l'activation des neurones d'une couche donnée via la résolution du calcul matriciel suivant :

$$y = f(Wx + b), \quad \text{où} \quad W = \begin{pmatrix} w_{0,0} & \cdots & w_{0,n} \\ \vdots & \ddots & \vdots \\ w_{n,0} & \cdots & w_{n,n} \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{pmatrix}. \quad (\text{II.29})$$

Pour une couche l donnée, chaque neurone i de cette couche effectue une somme pondérée de ses entrées x_1, x_2, \dots, x_n (qui sont les sorties des neurones de la couche précédente) avec la matrice de poids W qui est initialisée de manière aléatoire avec des valeurs proches de 0, ajoute un biais b_i , et passe le résultat à travers une fonction d'activation f . Pour chaque instance, l'algorithme alimente le réseau avec les entrées x et calcule la sortie de chaque neurone dans chaque couche consécutive. Ce processus se nomme la *forward propagation*. On obtient $y_i^{(l)}$ la valeur de sortie du neurone i dans une couche l :

$$y_i^{(l)} = f \left(\sum_{j=1}^n w_{ij}^{(l)} x_j^{(l-1)} + b_i^{(l)} \right) \quad (\text{II.30})$$

Le fonctionnement d'un tel réseau peut être schématisé de la manière suivante, illustrant une sortie finale unique, bien qu'il soit possible de concevoir des architectures avec plusieurs sorties :

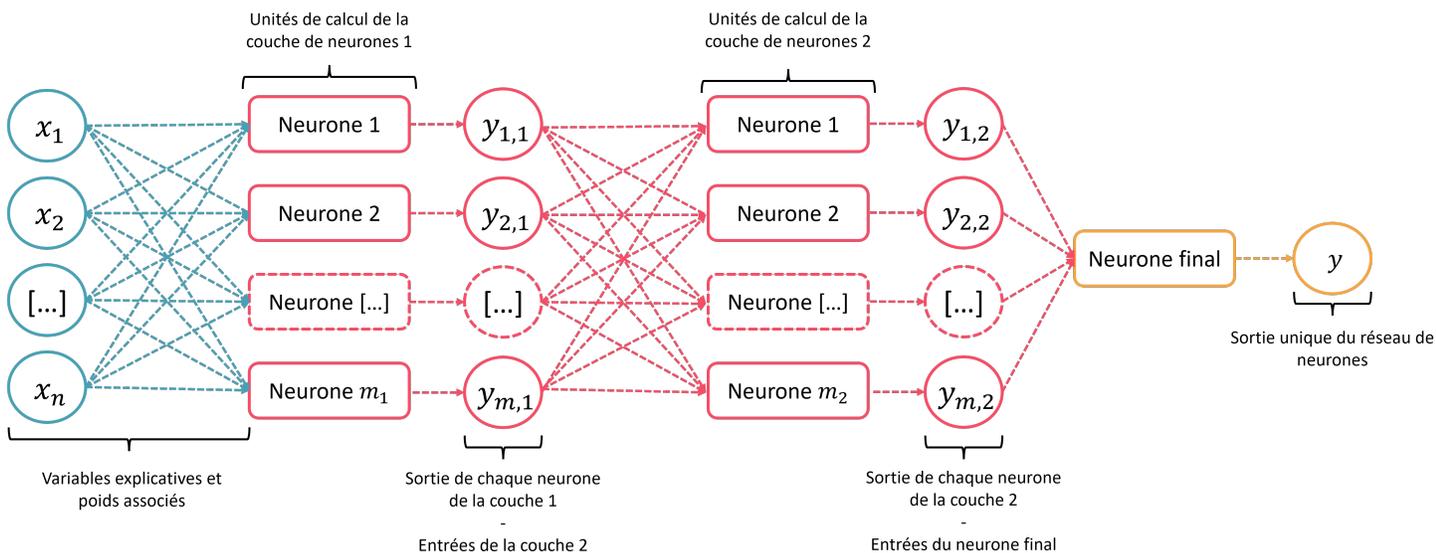


FIGURE II.7 – Représentation d'un réseau de neurones

L'optimisation des paramètres

La structure du réseau de neurones étant fixe, l'algorithme dans un réseau *fully-connected* ne peut que s'appuyer sur le biais et les poids afin d'optimiser ces résultats. À la fin du processus de *forward-propagation*, l'algorithme à l'aide d'une fonction de perte définie au préalable qu'on notera $L(\cdot)$ va mesurer l'erreur de la sortie y du réseau (c'est-à-dire la différence entre la sortie souhaitée et la sortie réelle du réseau). Notons θ l'ensemble des paramètres du réseau de neurones au sein d'un unique vecteur. Une fois la fonction de perte choisie, l'ajustement de la matrice de poids W se réduit au problème d'optimisation suivant :

$$\theta^* = \operatorname{argmin} L(y_{pred,\theta}, y)$$

où $y_{pred,\theta}$ est la prédiction réalisée par les réseaux de neurones et y la valeur réelle des observations.

Le problème d'ajustement des paramètres du réseau de neurones conduit alors à minimiser la fonction de perte. Étant donné que les poids influencent la perte, un réajustement de ces derniers s'impose mais ce processus peut s'appliquer à n'importe quel paramètre du vecteur θ . Par la suite, il sera étudié le cas de l'ajustement des poids, en considérant les biais comme nuls afin de simplifier les calculs, de façon à obtenir une combinaison qui minimise la fonction de perte.

Le processus de *back-propagation* commence alors, permettant de réajuster les erreurs des sorties. L'erreur est propagée vers l'arrière pendant la *back-propagation* de la couche de sortie à la couche d'entrée. Cette erreur est ensuite utilisée pour calculer le gradient de la fonction de coût par rapport à chaque poids comme décrit dans la figure II.8.

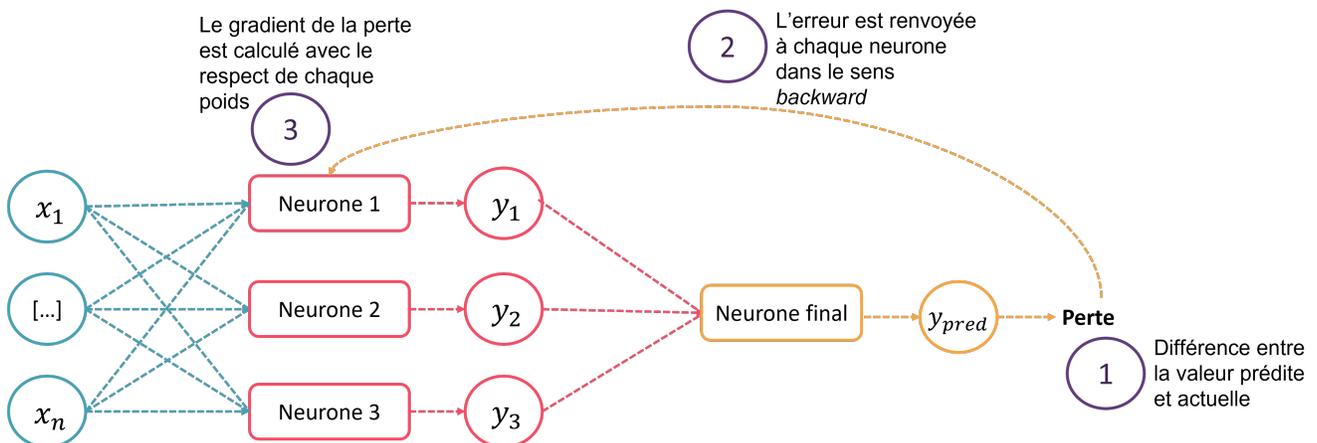


FIGURE II.8 – La *Backpropagation*

La *back-propagation* vise à calculer le gradient négatif de la fonction de perte. Ce gradient négatif aide à ajuster les poids. Le gradient de la fonction de perte, noté $\nabla_w L$, est un vecteur dont chaque

composante est la dérivée partielle de la fonction de perte par rapport à chaque poids, soit :

$$\nabla_w L = \left[\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \dots, \frac{\partial L}{\partial w_n} \right]$$

La *back-propagation* calcule ce gradient en utilisant la règle de la chaîne du calcul différentiel. Pour chaque poids w_i , la dérivée partielle $\frac{\partial L}{\partial w_i}$ est calculée en différenciant L par rapport à w_i , en traitant tous les autres poids comme constants.

Après avoir calculé le gradient pour chaque poids, ils sont alors ajustés à l'aide de la méthode de descente de gradient. Il s'agit d'un algorithme itératif d'optimisation qui permet de réduire la fonction de perte. Les poids sont ajustés à l'aide de la formule suivante :

$$w_i^{(nouveau)} = w_i^{(ancien)} - \alpha \frac{\partial L(y_{pred, \theta}, y)}{\partial w_i} \quad (\text{II.31})$$

où α est le *learning rate*.

Le *learning rate* est un paramètre qui détermine la taille du pas à chaque itération de la descente de gradient. La taille du pas joue un rôle important dans l'équilibre entre le temps d'optimisation et la précision. Un α faible correspond à une petite taille de pas, ce qui implique que le processus d'optimisation peut prendre beaucoup de temps. Si la taille de pas est excessivement grande, l'algorithme pourrait ne pas converger et ainsi manquer le minimum global.

Dans le cadre des méthodes de descente de gradient, les données sont entraînées par *batch*. Un *batch* fait référence à un sous-ensemble de l'ensemble de données d'apprentissage. En effet, il n'est généralement pas pratique, ni parfois même possible, de passer l'ensemble des données d'apprentissage à travers le réseau de neurones en une seule fois. Chaque *batch* de données est passé à travers du réseau, les gradients sont calculés et les poids du réseau sont mis à jour pour chaque *batch*.

Par exemple, si la base d'entraînement comporte $N = 200$ lignes et que la taille d'un *batch* est $T = 60$, alors la base est divisée aléatoirement en 3 *batches* de 60 observations et 1 *batch* de 20 observations.

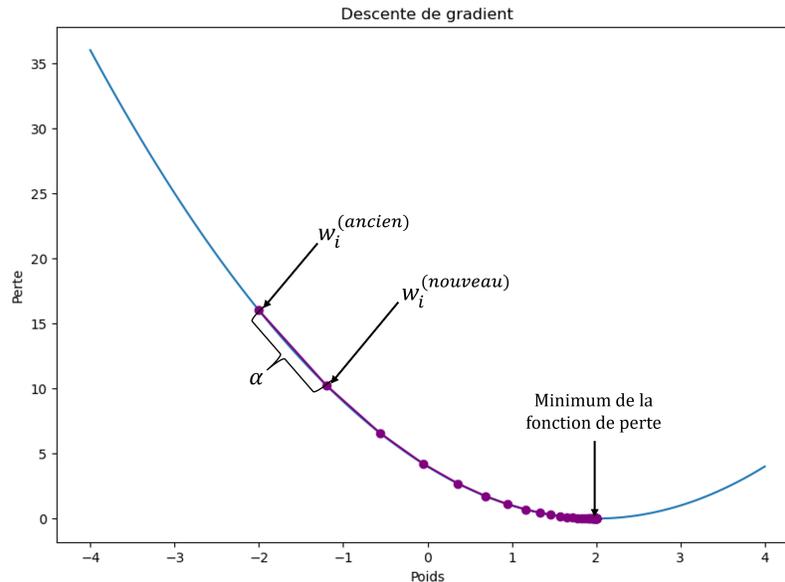


FIGURE II.9 – L’algorithme de descente de gradient

L’ajustement des poids consiste en de multiples itérations. À chaque itération, comme illustré ci-dessus par la figure II.9, l’algorithme descend d’un cran et calcule un nouveau poids à l’aide de la formule II.31 . En utilisant le poids initial, le gradient et le taux d’apprentissage, il est alors possible de déterminer les poids suivants et ainsi converger vers les poids optimaux permettant une minimisation de la fonction de perte.

Après avoir introduit le concept fondamental des réseaux de neurones et leurs fonctionnements, il est important de voir comment ils peuvent être utilisés pour résoudre des problèmes d’apprentissage par renforcement de manière plus efficace et robuste. En particulier, un algorithme d’apprentissage par renforcement, appelé *Deep Q-Learning* ou *Deep Q-Network*.

II.2.7 L’algorithme *Deep Q-Network*

D’après Mnih [2015], le *Deep Q-Network* (DQN) est un algorithme alliant le *Q-Learning* et les réseaux de neurones artificiels. Le RL est connu pour être instable, voire diverger, lorsqu’un réseau de neurones est utilisé afin de calculer une approximation de la fonction d’action-valeur Q . Les instabilités dans un réseau de neurones pour le RL sont résolues avec une nouvelle variante de l’apprentissage Q utilisant deux idées clés :

- Un *experience replay*
- Un réseau de neurones cible

Dans l’apprentissage par renforcement traditionnel, un agent apprend de nouvelles expériences en

continu. Cependant, ces expériences sont souvent fortement corrélées, ce qui peut mener à une instabilité et une divergence lors de l'apprentissage. Pour pallier à ce problème, l'*experience replay* stocke les expériences de l'agent (état, action, récompense, nouvel état) dans une mémoire de *replay* à chaque pas de temps. Ensuite, plutôt que d'apprendre en ligne à partir de l'expérience la plus récente, l'agent échantillonne un mini-lot, un *batch* d'expériences passées de manière aléatoire pour apprendre. Cette méthode de *replay* permet de s'affranchir des corrélations temporelles et d'exploiter les expériences passées, permettant aussi d'accélérer l'apprentissage.

L'une des principales différences entre le *Deep Q-Learning* et le *Q-Learning* réside dans l'implémentation de deux réseaux de neurones : un réseau cible et un réseau de prédiction. Ces réseaux ont la même architecture mais des paramètres différents. Toutes les C itérations (un hyperparamètre), les paramètres du réseau de prédiction θ sont copiés dans le réseau cible θ^- .

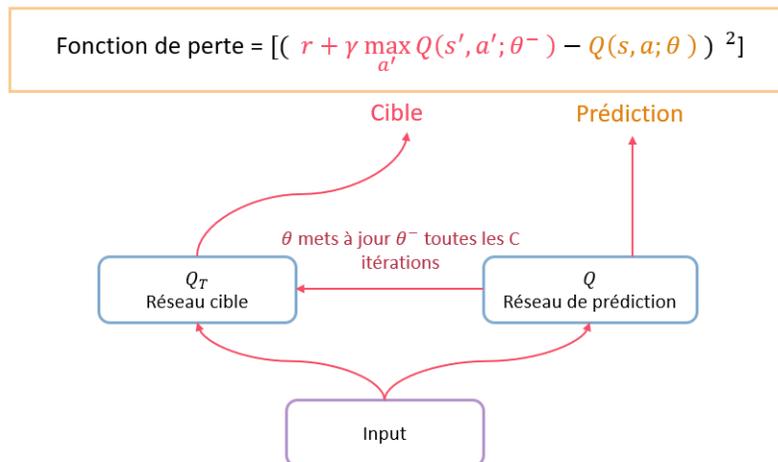


FIGURE II.10 – L'algorithme du *Deep Q-Learning*

Pour formaliser l'*Experience Replay*, on considère un ensemble d'expériences $D = (s_t, a_t, r_t, s_{t+1})$, où s_t est l'état à l'instant t , a_t est l'action prise, r_t est la récompense obtenue et s_{t+1} est l'état suivant. À chaque pas de temps t , une expérience (s_t, a_t, r_t, s_{t+1}) est stockée dans D . Pendant la phase d'apprentissage, un *batch* d'expériences est sélectionné aléatoirement à partir de l'ensemble D pour la mise à jour des paramètres du réseau de neurones. Ce *batch* est ensuite introduit dans la case "Input" du schéma II.10. Les réseaux de neurones cibles et de prédiction produisent respectivement les sorties Q_T et Q . Ces deux sorties permettent de calculer la fonction de perte introduite ci-dessus. Enfin, les paramètres θ sont mis à jour à l'aide de la méthode de descente de gradient afin de minimiser la fonction de perte tandis que les paramètres θ^- sont mis à jour toutes les C itérations avec les paramètres du réseau Q .

II.2.8 Optimisation à l'aide du gradient de la politique

Les méthodes présentées ci-dessus visent à apprendre la fonction d'action-valeur. Les méthodes de gradient de politique optimisent directement la politique à l'aide d'une fonction paramétrée en fonction de θ , $\pi_\theta(a|s)$ au travers d'un réseau de neurones. Contrairement aux méthodes d'optimisation de la fonction d'action-valeur, au lieu de minimiser la fonction de perte, ces méthodes ont pour objectif de maximiser une fonction de récompense notée $J(\theta)$. La fonction de récompense, ou d'objectif, est définie comme l'espérance de la somme des récompenses futures à partir de l'état initial, sous la politique π_θ . Mathématiquement, on la définit comme suit :

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [G(\tau)] = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=1}^T \gamma^t r_t \right] \quad (\text{II.32})$$

où $\mathbb{E}_{\tau \sim \pi_\theta} [G(\tau)]$ est le rendement d'une trajectoire introduit précédemment. Dans un espace continu :

$$J(\theta) = \int_{\mathcal{S}} \rho_\pi(s) \int_{\mathcal{A}} \pi_\theta(s, a) Q_\pi(s, a) da ds \quad (\text{II.33})$$

où ρ_π est la distribution stationnaire de la chaîne de Markov pour π_θ . Par souci de simplicité, le paramètre θ est omis pour la politique π_θ lorsque la politique est présente dans l'indice d'autres fonctions, par exemple, ρ_π et Q_π devraient être ρ_{π_θ} et Q_{π_θ} .

La méthode du gradient de la politique consiste à ajuster les paramètres θ pour maximiser $J(\theta)$ en utilisant la descente de gradient afin de produire un rendement maximal. Le gradient de la politique est souvent calculé en utilisant le théorème de *Policy Gradient*.

II.2.9 Le théorème du gradient de la politique

Le gradient de la politique est calculé en fonction des paramètres du réseau de neurones et est utilisé pour mettre à jour la politique. La descente de gradient est formulée de la manière suivante :

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta_t)$$

où $\nabla_\theta J(\theta_t)$ est le gradient de la politique.

Cependant, dans la pratique, il n'est pas possible de déterminer le véritable gradient de la fonction objectif car cela nécessiterait l'évaluation de la probabilité de chaque trajectoire possible, ce qui est extrêmement coûteux en termes de calculs. L'objectif est donc de calculer une estimation du gradient à partir d'une estimation basée sur des échantillons (collecte de certaines trajectoires). De plus, cette fonction n'est pas obligatoirement différentiable du fait que les probabilités de chaque trajectoire ne sont pas forcément connues.

Pour cela, Sutton et Barto [2017] introduisent le "*Policy Gradient Theorem*" qui fonctionne même lorsque $J(\theta)$ n'est pas différentiable, la démonstration d'un tel résultat est disponible en Annexe A.5 :

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\pi_{\theta}} [\nabla_{\theta} \ln \pi_{\theta}(a | s) Q_{\pi}(s, a)] \quad (\text{II.34})$$

Enfin, les paramètres de la politique sont mis à jour comme suit :

$$\theta_{t+1} = \theta_t + \alpha \nabla_{\theta} \ln \pi_{\theta}(a_t | s_t) Q_{\pi}(s_t, a_t)$$

Le théorème du gradient de la politique est une notion centrale dans beaucoup de catégories d'algorithmes de RL notamment les méthodes *Actor-Critic*. Ces algorithmes, conçus pour l'apprentissage par renforcement, constituent une harmonieuse fusion des approches dites *value-based* et *policy-based*.

II.2.10 Les méthodes *Actor-Critic*

Les méthodes *Actor-Critic* (Weng [2018]) utilisent un réseau de neurones pour estimer le gradient de la politique comme présenté dans la sous-section II.2.8 avec une seconde fonction qui estime la fonction d'action-valeur $Q^{\pi}(s, a)$. Ce modèle utilise un agent qui prend des décisions (*Actor*) et un critique (*Critic*) indiquant si les actions entreprises par l'agent sont bonnes ou non. À chaque itération, les paramètres de l'agent et du critique sont mis à jour.

C'est l'idée principale d'un algorithme *Actor-Critic*. Plus formellement l'algorithme estime deux fonctions (deux réseaux de neurones) :

- L'*Actor*, une politique qui contrôle les actions de l'agent de paramètre $\theta : \pi_{\theta}(s)$
- Le *Critic*, une fonction d'action-valeur de paramètre w permettant d'aider à la mise à jour de la politique en critiquant la qualité de l'action : $\hat{Q}_w(s, a)$

L'algorithme fonctionne comme illustré dans la figure II.11 ci-dessous :

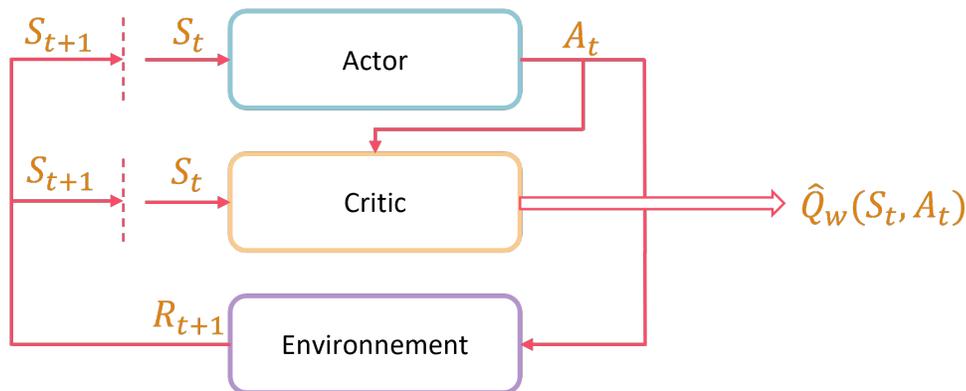


FIGURE II.11 – L'algorithme *Actor-Critic*

Le processus de l'algorithme, à chaque pas de temps t , utilise en entrée l'état actuel s_t de l'environnement et l'introduit au travers de l'*Actor* et du *Critic*. L'*Actor*, produit en suivant la politique suggérée par la figure II.11 une action a_t . Cette action a pour conséquence de générer un nouvel état s_{t+1} et une récompense r_{t+1} .

Le *Critic* en tenant compte de l'action et de l'état, calcule une action-valeur $\hat{Q}_w(s, a)$. Cette valeur permet par la suite de mettre à jour les paramètres du réseau de l'*Actor* en reprenant le théorème II.34 du gradient de la politique et l'estimation $\hat{Q}_w(s, a)$:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} (\log \pi_{\theta}(s, a)) \hat{Q}_w(s, a) \quad (\text{II.35})$$

Grâce à cette mise à jour des paramètres, l'*Actor* prend une nouvelle action a_{t+1} étant donné le nouvel état s_{t+1} .

Suite à cela, le *Critic* ajuste ses paramètres :

$$w \leftarrow w + \beta (r + \gamma \hat{Q}_w(s', a') - \hat{Q}_w(s, a)) \nabla_w \hat{Q}_w(s, a) \quad (\text{II.36})$$

où s' et a' représentent le nouvel état et la nouvelle action, β le *learning rate* propre au *Critic*.

Dans les méthodes évoquées ci-dessus, la politique π_{θ} est modélisée comme une probabilité de distribution des actions possibles dans un état donné, ce qui implique que la politique est stochastique. La version déterministe du gradient de la politique (DPG) modélise la politique de manière déterministe μ afin de traiter plus efficacement les espaces d'actions continus.

II.2.11 Gradient déterministe de la politique

Selon Silver [2014], une version déterministe du "*policy gradient theorem*" est présentée permettant de traiter efficacement les espaces d'actions continus en modélisant la politique comme une décision déterministe avec $a_t = \mu_{\theta}(s_t)$.

La fonction objective à optimiser est la suivante :

$$J(\theta) = \int_{\mathcal{S}} \rho^{\mu}(s) Q_{\mu}(s, a) ds \quad (\text{II.37})$$

où ρ^{μ} représente la distribution stationnaire de la chaîne de Markov pour μ_{θ} . Il est à noter que la politique étant déterministe, seul $Q^{\mu}(s, a)$ est nécessaire, plutôt que $\sum_a \pi(a|s) Q^{\pi}(s, a)$, pour estimer la récompense d'un état donné s .

La formule du gradient de politique déterministe est donnée par :

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \int_{\mathcal{S}} \rho^{\mu}(s) \nabla_a Q_{\mu}(s, a) \nabla_{\theta} \mu_{\theta}(s) \Big|_{a=\mu_{\theta}(s)} ds \\ &= \mathbb{E}_{\rho^{\mu}} \left[\nabla_a Q_{\mu}(s, a) \nabla_{\theta} \mu_{\theta}(s) \Big|_{a=\mu_{\theta}(s)} \right]\end{aligned}\tag{II.38}$$

La formule du gradient de politique déterministe ainsi que les concepts développés du *Deep Q-Network* (II.2.7) et les méthodes *Actor-Critic* (II.2.10) constituent le fondement de l'algorithme connu sous le nom de *Deep Deterministic Policy Gradient* qui est au cœur de cette étude.

Chapitre III

Application de la méthode DDPG : étude et intégration dans le modèle ALM

III.1 L'algorithme *Deep Deterministic Policy Gradient*

L'un des principaux objectifs du RL est de résoudre des problèmes complexes de haute dimension. Le DQN ne peut traiter que des espaces d'action discrets et de faible dimension, alors qu'il peut résoudre des problèmes avec des espaces d'observation de haute dimension. D'après Timothy [2016], l'algorithme *Deep Deterministic Policy Gradient* (DDPG) est une technique d'apprentissage par renforcement spécialement conçue pour gérer les situations d'action continue. Le DDPG est un algorithme qui tire parti de la structure des algorithmes *actor-critic*, du DQN, et du *Deterministic Policy Gradient* pour fonctionner efficacement dans les environnements d'action continue. Le terme "déterministe" signifie qu'il n'y a pas d'aléa ou de variabilité dans les résultats d'un système déterministe, ce qui contraste avec la politique stochastique. Une politique déterministe signifie que l'environnement RL produira toujours la même action pour un état donné.

Le DDPG est un algorithme qui reprend donc les concepts vus en II.2.7 sur le DQN, notamment l'*experience replay* et le réseau de neurones cible. DDPG utilise également deux ensembles de réseaux neuronaux *Actor-Critic* pour l'approximation des fonctions. Les deux ensembles se composent d'un réseau *Actor* et d'un réseau *Critic*, chacun ayant la même structure et la même paramétrisation.

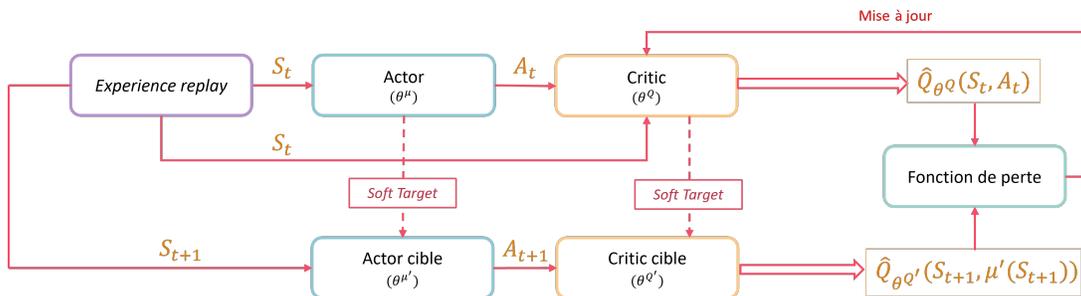


FIGURE III.1 – L'algorithme *Deep Deterministic Policy Gradient*

La mise à jour *soft target* comme le suggère la figure ci-dessus III.1 est utilisée dans le DDPG pour mettre à jour lentement les poids du réseau cible au lieu de copier directement les poids du réseau *Actor-Critic*. Cette méthode diffère du DQN car la mise à jour s'effectue à tous les pas de temps (contrairement au DQN qui effectue une mise à jour toutes les C itérations). L'approche *soft target* augmente considérablement la stabilité de l'apprentissage, et les mises à jour sont effectuées comme suit :

$$\begin{aligned}\theta^{Q'} &\leftarrow \tau\theta^Q + (1 - \tau)\theta^{Q'} \\ \theta^{\mu'} &\leftarrow \tau\theta^\mu + (1 - \tau)\theta^{\mu'}\end{aligned}\tag{III.1}$$

Une petite partie des poids du réseau *Actor-Critic* est envoyée au réseau cible à chaque pas de temps.

Le taux de mise à jour de la cible τ , un hyperparamètre, contrôle ce pourcentage. On considère généralement un $\tau \ll 1$.

L'algorithme emploie dans le même objectif que le DQN l'*experience replay* pour stocker les transitions et les récompenses $D = (s_t, a_t, r_t, s_{t+1})$ cela permet donc de minimiser les corrélations entre les échantillons et en sauvegardant les expériences antérieures, l'algorithme est en mesure de tirer des enseignements d'une variété de trajectoires différentes.

L'algorithme DDPG implique un réseau *Actor-Critic* avec une politique μ suivie par l'*Actor*, ainsi qu'un *Critic* évaluant les actions de l'*Actor* utilisant une fonction d'action-valeur. Le DDPG possède aussi un réseau cible *Actor-Critic*, avec μ' et Q' respectivement. Ces deux réseaux peuvent se décliner au travers des entités suivantes :

1. ***Actor* (θ^μ)** : Le réseau de neurones de l'*Actor* est de paramètre θ , prenant une observation s_t en tant qu'entrée du *experience replay* et renvoie l'action correspondante a_t qui maximise la récompense à long terme suivant la politique μ .

Comme discuté précédemment la politique μ est déterministe ce qui implique donc qu'à moins qu'il n'y ait suffisamment de bruit dans l'environnement, il est très difficile de garantir une exploration suffisante en raison du caractère déterminant de la politique. Pour cela, le DDPG ajoute du bruit d'exploration aux actions sélectionnées par l'*Actor* afin de relever le défi de l'équilibre l'exploitation-exploration d'espaces d'action continus.

Le processus Ornstein-Uhlenbeck (III.5) est une technique généralement utilisée dans l'algorithme DDPG pour introduire une exploration continue dans l'espace d'action. Une nouvelle politique d'exploration, μ_{new} , est construite en ajoutant du bruit \mathcal{N} échantillonné à partir du processus d'Ornstein-Uhlenbeck.

$$\mu_{new}(s_t) = \mu(s_t | \theta_t^\mu) + \mathcal{N} \quad (\text{III.2})$$

Mathématiquement, le processus Ornstein-Uhlenbeck est défini par l'équation différentielle stochastique suivante :

$$dx_t = -\theta(x_t - \mu)dt + \sigma dW_t, \quad (\text{III.3})$$

où :

- x_t est la position de la particule à l'instant t ,
- θ est un terme de retour vers la moyenne,
- μ est la valeur de la moyenne à long terme,
- σ est la volatilité,
- dW_t est un mouvement brownien (un processus stochastique qui génère un bruit blanc gaussien).

En ajoutant ce bruit, donné par l'équation (III.5), au choix de l'action dans l'algorithme DDPG, on encourage l'exploration continue, ce qui peut aider l'algorithme à apprendre une meilleure politique et ainsi améliorer les performances de l'algorithme.

Ainsi l'action transmise par l'Actor au Critic sera de la forme suivante : $a_t = a_t + \mathcal{N}$

Pour mettre à jour le réseau de l'Actor (θ^μ), la variante du gradient déterministe de la politique est employée. Cela signifie que le gradient de la politique peut être calculé plus simplement, en utilisant la formule du gradient de politique déterministe introduit en II.38. Ainsi les mises à jour de la politique (ou l'acteur) peuvent être réécrites comme suit :

$$\theta^\mu \leftarrow \theta^\mu + \alpha \nabla_{\theta^\mu} J(\theta)$$

avec :

$$\nabla_{\theta^\mu} J(\theta) = \mathbb{E}[\nabla_a \hat{Q}_{\theta^Q}(s, a) \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{a=\mu(s|\theta^\mu)}]$$

Ici, $\mu(s|\theta^\mu)$ est la politique déterministe qui choisit une action en fonction de l'état s , et θ^μ sont les paramètres de l'Actor. La mise à jour est effectuée pour maximiser la fonction d'action-valeur $\hat{Q}_{\theta^Q}(s, a)$ fournie par le Critic(θ^Q).

2. **Actor cible** ($\theta^{\mu'}$) : Le réseau de l'Actor cible est une copie du réseau de l'Actor, mais ses paramètres sont mis à jour plus lentement.

Les paramètres du réseau d'acteur cible $\theta^{\mu'}$ sont mis à jour pour se rapprocher lentement des paramètres du réseau d'acteur comme introduit selon l'équation III.1 :

$$\theta^{\mu'} \leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'}$$

Les réseaux cibles sont utilisés pour accroître la stabilité du processus d'apprentissage. Les cibles restent relativement constantes, au moins pendant une courte période, ce qui favorise la convergence de l'algorithme.

3. **Critic** (θ^Q) : Le réseau de neurones Critic avec les paramètres θ^Q prend l'observation s_t provenant de l'expérience replay et l'action a_t qui est la sortie du réseau de neurones Actor (θ^μ) après l'ajout du bruit et produit la valeur attendue correspondante $\hat{Q}_{\theta^Q}(s, a)$ fonction d'action-valeur à l'aide de l'équation de Bellman. Dans le cadre de l'algorithme DDPG, le Critic est en fait très similaire à l'approche utilisée dans le DQN. En effet, à l'instar du DQN, le Critic et le Critic cible cherchent à minimiser une fonction de perte :

$$L = \frac{1}{N} \sum_i (y_i - \hat{Q}_{\theta^Q}(s_i, a_i))^2$$

avec y_i la sortie du Critic cible qui sera détaillée par la suite.

Le *Critic* utilise ensuite cette erreur pour mettre à jour les poids de son réseau de neurones par *back-propagation* (II.31).

4. **Critic cible** ($\theta^{Q'}$) : Le réseau du *Critic* cible, noté Q' , a pour rôle d'estimer la valeur de la politique de l'*Actor* cible μ' dans le nouvel état s_{t+1} . C'est-à-dire, qu'il est utilisé pour calculer la sortie du *Critic* cible $\hat{Q}_{\theta^{Q'}}(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'}))$, qui sera utilisée ensuite pour calculer y_i de la fonction de perte :

$$y_i = r(s_i, a_i) + \gamma \hat{Q}_{\theta^{Q'}}(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'}))$$

Le réseau du *Critic* cible est mis à jour à partir du réseau du *Critic*, de manière analogue avec l'*Actor* cible et l'*Actor* afin d'assurer la stabilité de l'apprentissage :

$$\theta^{Q'} \leftarrow \tau \theta^Q + (1 - \tau) \theta^{Q'}$$

Le DDPG est une méthode puissante qui allie de nombreux concepts de RL, capable de résoudre des tâches complexes avec des espaces d'actions continus. Cependant, comme tout algorithme d'apprentissage par renforcement, il requiert une attention particulière lors de sa paramétrisation afin d'obtenir des résultats convenables. L'algorithme DDPG est la méthode retenue pour le cadre de cette étude. Néanmoins, sa mise en œuvre nécessite une paramétrisation méticuleuse pour exploiter des résultats convenables. C'est dans ce contexte qu'il s'avère nécessaire de décrire la méthodologie de travail.

III.2 Méthode de travail adoptée

Le DDPG est un modèle qui requiert, comme tout algorithme de RL ou plus généralement de *Machine Learning*, une phase d'entraînement. Dans le cadre de cette étude, l'algorithme est employé dans deux univers neutre et réel présentés respectivement en I.5.1 et I.5.2. L'objectif de cet algorithme est d'optimiser différentes métriques au travers de stratégies d'allocation d'actifs. Le choix de ces métriques dépend de l'univers de projection. Pour l'univers risque neutre deux métriques d'optimisation sont retenues : le $SCR_{marché}$ et la PVFP. Ces deux métriques sont complémentaires afin de permettre à la fois de respecter les exigences réglementaires ainsi qu'optimiser le résultat financier de l'assureur. En se concentrant sur le $SCR_{marché}$ au lieu du SCR_{global} , l'étude vise à isoler et à examiner l'influence directe des décisions d'allocation d'actifs sur la partie du bilan la plus sensible aux fluctuations du marché. Pour l'univers risque réel, il s'agit du TRA et de la richesse latente. Le TRA se concentre sur le rendement des actifs, tandis que la richesse latente offre une vision plus holistique du passif et de l'actif de l'assureur. Cette dualité d'indicateurs répond aux enjeux de l'assureur en matière de rentabilité financière et de gestion des fonds propres.

Une fois les métriques choisies, l'étape suivante consiste à mettre en œuvre les modèles de DDPG dans chacun des univers de risque. Les algorithmes sont soumis à une phase d'entraînement afin de calibrer les paramètres des modèles pour optimiser les métriques sélectionnées. Les objectifs d'optimisation sont décrits par le tableau III.1.

Univers de Risque	Métrique	Objectif
Risque Neutre	PVFP	Maximiser
	$SCR_{marché}$	Minimiser
Risque Réel	Richesse Latente	Maximiser
	TRA	Maximiser

TABLE III.1 – Objectifs d'optimisation pour les univers de risque neutre et réel

Chaque modèle de DDPG, une fois entraîné, fournit une allocation d'actifs optimale en fonction des métriques ciblées. Ces stratégies de DDPG seront ensuite mises en comparaison avec la stratégie actuelle, qui est basée sur une approche *Fixed-Mix*. La méthodologie est synthétisée par le schéma ci-dessous :

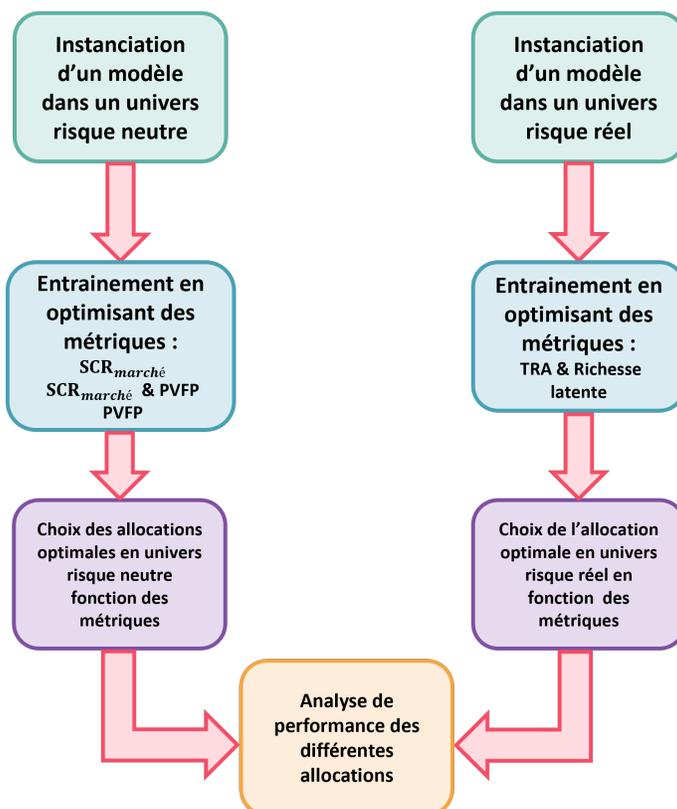


FIGURE III.2 – Description de la méthodologie

En résumé, quatre allocations d'actifs sont proposées par le modèle. Pour l'univers risque neutre, trois modèles distincts sont utilisés : le premier optimise le SCR de marché, le deuxième se concentre sur la PVFP, et le troisième vise à optimiser à la fois le SCR de marché et la PVFP. L'utilisation de ces trois modèles en univers risque neutre permet également de réaliser des études de sensibilité et d'analyser le comportement de l'algorithme sous différents scénarios d'optimisation. L'évaluation de performance des modèles en univers risque neutre permet d'avoir des éléments de comparaison notamment au travers des bilans Solvabilité II. De plus, un modèle est développé dans un univers risque réel, qui optimise à la fois le TRA et la richesse latente en fin de projection afin de pouvoir étudier le comportement de l'algorithme sous deux univers de projection. Cette comparaison permet d'obtenir une vision à la fois économique et réglementaire des différentes stratégies.

Après avoir établi les objectifs d'optimisation pour chaque univers de risque, l'étape suivante consiste en la détermination des paramètres du modèle DDPG, communément appelé phase d'entraînement. Cette étape est fondamentale car elle conditionne la qualité des allocations d'actifs optimales que les modèles sont censés générer.

III.3 Phase d'entraînement : approche et paramétrage

Pour entraîner le modèle de DDPG, il est nécessaire d'explicitier les différentes composantes du modèle introduit en partie II.2.1. Cela inclut la définition empirique de l'espace d'états, de l'espace d'actions, de la fonction de récompense.

III.3.1 Configuration de l'espace d'états, d'actions et récompense

Action

Dans le contexte de cette étude, l'action prise par l'agent est assimilée à la décision d'allocation d'actifs à chaque pas de temps. L'agent évalue l'état actuel du portefeuille et choisit une action afin de répartir le capital sur l'ensemble des classes d'actifs (le résidu de capital non attribué va dans le cash) pour optimiser les métriques cibles, que ce soit le $SCR_{marché}$ et la PVFP dans un univers de risque neutre, ou le TRA et la Richesse Latente dans un univers de risque réel.

Pour adopter une stratégie d'allocation d'actifs cohérente, il est nécessaire d'adapter la réponse de l'agent par le biais de la fonction d'activation. En effet, l'objectif du modèle est de fournir un pourcentage pour les différentes classes d'actifs : actions, immobilier, obligations (état et entreprise) et cash. La fonction d'activation *softmax* permet d'obtenir de tels pourcentages :

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^N e^{x_j}} \quad (\text{III.4})$$

où :

- x_i la sortie du réseau de neurones pour la classe d'actif i
- N le nombre total de classes d'actifs

Cette fonction transforme les sorties du réseau de neurones en une distribution de probabilités, où la somme des sorties est égale à 1.

Le modèle de DDPG est également soumis à des contraintes d'allocation qui sont représentatives des conditions de marché. Ces contraintes ont été construites par l'observation de différentes stratégies d'allocations d'actifs de différents acteurs de marché. De plus, afin de faire face à des problématiques d'illiquidités du marché immobilier, il est considéré que la part d'immobilier ne peut varier que de $\pm 2\%$ par pas de temps. De même, le modèle impose que la part des actions ne puisse varier que de $\pm 8\%$ et celle des obligations de $\pm 4\%$ par pas de temps. Enfin, pour maintenir un certain niveau de liquidité pour l'assureur, le modèle a pour contrainte de garder environ 3 % du capital en monétaire. Ces contraintes sont récapitulées comme suit :

Classe d'actif	Limite inférieure	Limite supérieure
Actions	5%	20%
Immobilier	5%	18%
Obligations	65%	85%

TABLE III.2 – Limites d'allocation pour les classes d'actifs

Classe d'actif	Actions	Immobilier	Obligations
Variation par pas de temps	±8%	±2%	±4%

TABLE III.3 – Limites de variation par pas de temps pour les classes d'actifs

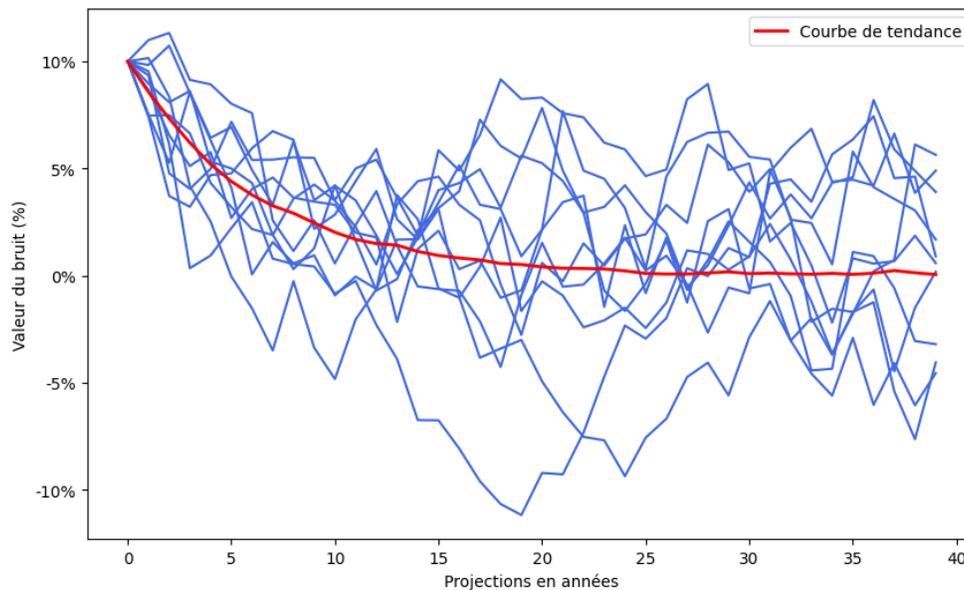
Par la suite, afin de pouvoir remédier à la problématique "exploitation-exploration", du bruit est ajouté dans l'espace d'action afin de permettre à l'agent d'explorer de nouvelles solutions. Le processus d'Ornstein-Uhlenbeck III.5 est défini afin de répondre à cette problématique. En effet, contrairement au bruit gaussien, il est corrélé dans le temps, ce qui signifie qu'il conserve une certaine mémoire de l'état précédent et permet d'éviter des changements brusques d'actions, ce qui est souhaitable dans l'étude et plus représentatif de la réalité.

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t \quad (\text{III.5})$$

où :

- θ est le paramètre de retour à la moyenne.
- μ est la moyenne de long terme vers laquelle le processus tend.
- X_t représente la valeur du processus à l'instant t .
- σ est le terme de volatilité.
- W_t est un processus de Wiener.

Le modèle requiert un paramétrage robuste afin de pouvoir produire des résultats adaptés à l'étude. Dans la littérature (Timothy [2016]), il est recommandé d'introduire un niveau élevé de bruit au début de l'entraînement afin d'explorer efficacement l'ensemble des actions possible, pour cela un bruit initial autour de 10 % est retenu. Dans le processus d'entraînement, cela signifie que, sur une décision d'investissement prise par l'agent, le bruit est ajouté au pourcentage d'allocation alloué à chaque d'actifs. À mesure que le nombre de simulations augmente au cours de l'entraînement, ce niveau de bruit est réduit de manière progressive jusqu'à atteindre une valeur proche de zéro. Cette stratégie permet une exploration de l'espace des actions et d'éviter de se restreindre à des optimums locaux.



Paramètre	Valeur
μ (Moyenne à long terme)	0
θ (Taux de retour à la moyenne)	0.15
σ (Volatilité du processus)	0.02
Valeur initiale	0.1

FIGURE III.3 – Processus d'Ornstein-Uhlenbeck sur 40 ans

Le graphique ainsi que le tableau III.3 illustre les différentes valeurs prises par le processus sur différentes trajectoires. La courbe de tendance représente la moyenne de 1000 simulations permettant de mettre en évidence le comportement général du processus.

État

L'espace d'état est un élément crucial dans tout modèle de renforcement, car il définit les informations disponibles pour l'agent afin de prendre des décisions pertinentes. Dans le modèle DDPG pour la gestion d'actifs, l'espace d'état dépend de l'univers de projection, il est constitué des éléments suivants :

— Monde réel

$$\text{État}_{\text{réel}} = \left(\text{Richesse Latente, TRA, Indice}_{\text{action}}, \right. \\ \left. \text{Indice}_{\text{immobilier}}, \text{Indice}_{\text{obligation}}, \right. \\ \left. \text{Allocation}_{\text{action}}, \text{Allocation}_{\text{immobilier}}, \text{Allocation}_{\text{obligation}} \right)$$

— Monde neutre

$$\text{État}_{\text{neutre}} = \left(\text{SCR}_{\text{marché}}, \text{PVFP}, \text{Indice}_{\text{action}}, \right. \\ \left. \text{Indice}_{\text{immobilier}}, \text{Indice}_{\text{obligation}}, \right. \\ \left. \text{Allocation}_{\text{action}}, \text{Allocation}_{\text{immobilier}}, \text{Allocation}_{\text{obligation}} \right)$$

L'agent prend en compte la valeur des métriques en fin de projection ainsi que, pour chaque indice, 40 valeurs correspondantes aux données extraites du GSE et les 40 années d'allocations d'actifs. L'état évolue dynamiquement en fonction des choix effectués au fur et à mesure de la projection. Après avoir opéré une action, chacune des valeurs constituant l'état est mise à jour pour refléter les nouvelles conditions du portefeuille et du modèle. De plus, les états sont réinitialisés au début de chaque nouvelle simulation du modèle ALM. Cela a pour but d'obtenir des résultats comparables entre les simulations qui permettent de mettre à jour la récompense de l'agent.

Récompense

La fonction de récompense est conçue pour guider l'agent à prendre des décisions qui sont à la fois optimales et réalisables. Elle est composée de plusieurs composantes, chacune ayant un objectif spécifique.

- **Pénalités pour dépassement des limites** : Pour chaque classe d'actifs (actions, immobilier, obligations) si l'action prise par l'agent dépasse ces limites III.2, une pénalité est appliquée à la fonction de récompense.

$$\text{Pénalité}_{\text{limites}} = -p \times |\text{Action}_i - \text{Limite}_{\text{actif}}|$$

où p est une constante fixe, Action_i est l'allocation pour la classe d'actifs i , et $\text{Limite}_{\text{actif}}$ est la limite supérieure ou inférieure pour cette classe d'actifs. La constante p est déterminée en se basant sur les recommandations d'experts.

- **Pénalité pour variation d'un actif** : Si l'action prise par l'agent modifie la proportion d'un actif d'une année sur l'autre (actions, immobilier, obligations) au-delà des limites définies dans le tableau III.3, alors une pénalité est appliquée.

$$\text{Pénalité}_{\text{variation}} = -n \times |\text{Action}_{\text{actif},t} - \text{Action}_{\text{actif},t-1}|$$

où n est une constante fixe, $\text{Action}_{\text{actif},t}$ l'allocation d'un actif au temps de projection t .

- **Pénalité pour la somme des actions** : Pour garantir qu'une partie de l'allocation est attribuée au cash, une contrainte incite le modèle à fournir une allocation proche de 97 % afin de pouvoir affecter le résidu en monétaire. Réserver 3 % des actifs en cash assure à l'assureur une marge de manœuvre financière, essentielle pour gérer les exigences de liquidité immédiates.

$$\text{Pénalité}_{\text{somme}} = -j \times \left| \sum_{i=1}^n \text{Action}_i - 0.97 \right|$$

où j est une constante fixe.

- **Récompense pour amélioration d'une métrique** :

La récompense basée sur l'amélioration d'une métrique spécifique encourage l'agent à rechercher des actions qui maximisent ou minimisent (selon l'objectif) la valeur de cette métrique. La valeur de la métrique à la fin de chaque simulation ($t = 40$ ans) est comparée avec l'historique de cette métrique des simulations précédentes.

$$\text{Récompense}_{\text{métriques}} = m \times |\text{Métrique}(t, i) - \text{Métrique}(0, i)|, \quad \text{si } \text{Métrique}(t, i) > \max_{j < i} \text{Métrique}(t, j)$$

ou/et

$$\text{Récompense}_{\text{SCR}_{\text{marché}}} = m \times |\text{SCR}_{\text{marché}}(t, i) - \text{SCR}_{\text{marché}}(0, i)|, \quad \text{si } \text{SCR}_{\text{marché}}(t, i) < \min_{j < i} \text{SCR}_{\text{marché}}(t, j)$$

où m est une constante fixe, $\text{Métriques}(t, i)$ représente la valeur de la métrique pendant la simulation i à l'instant t et appartient à l'ensemble {TRA, Richesse Latente, PVFP}.

La fonction de récompense totale est donc la somme de ces différentes composantes :

$$\text{Récompense}_{\text{total}} = \text{Récompense}_{\text{métriques}} + \text{Récompense}_{\text{SCR}_{\text{marché}}} + \text{Pénalité}_{\text{limites}} + \text{Pénalité}_{\text{variation}} + \text{Pénalité}_{\text{somme}}$$

Il est important de préciser que les contraintes définies dans la fonction de récompense ne sont pas absolues. Par exemple, pour la pénalité pour dépassement des limites, l'agent n'est pas pénalisé s'il respecte ces contraintes, mais il n'est pas strictement empêché de le faire. Cette stratégie a pour but de permettre à l'algorithme d'explorer plus librement l'espace d'actions.

III.3.2 Calibrage du modèle

Le processus de calibrage du modèle est une étape nécessaire dans l'utilisation de modèle de *reinforcement learning*. En effet, ce processus permet de s'assurer de la robustesse et de la pertinence des sorties du modèle. Le calibrage reprend le principe du schéma II.3 d'introduction de l'agent dans son environnement. L'environnement, qui est dans le cas de cette étude le modèle ALM, fournit à l'agent un état en fonction du monde de projection. Ensuite, l'agent effectue une action qui est le choix d'une allocation d'actifs à l'instant t puis observe la récompense afin de pouvoir s'entraîner.

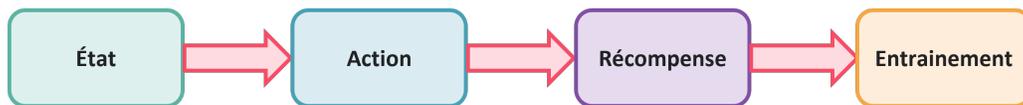


FIGURE III.4 – Description de la méthode d'entraînement du modèle

Le schéma III.4 illustre le processus de l'entraînement de l'agent pour une année de projection. Pour entraîner le modèle efficacement, le processus ci-dessus est itéré sur plusieurs épisodes. Chaque épisode d'entraînement est composé de 1000 simulations, et chaque simulation comprend 40 années de projections. Le modèle est entraîné sur un total de 200 épisodes, ce qui équivaut à 200 000 trajectoires uniques, chacune avec 40 années de projections. Cette approche permet d'assurer que le modèle est suffisamment robuste et capable de généraliser face à de nouvelles situations.

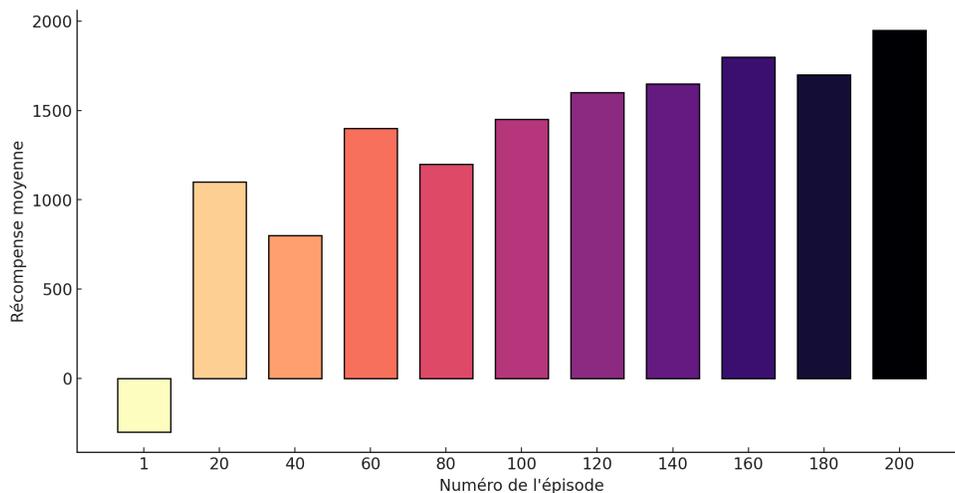


FIGURE III.5 – Exemple de l'évolution de la récompense moyenne pendant la calibration du modèle

Le graphique III.5 illustre le processus de calibrage du modèle risque neutre optimisant le $SCR_{marché}$ et la PVFP au travers du mécanisme de la récompense. La tendance générale des récompenses semble être croissante, indiquant que l'algorithme améliore sa performance au fil des épisodes. Des fluctuations sont observées, ce qui est typique dans un processus d'apprentissage par renforcement en raison

du compromis entre exploration et exploitation. En effet, l'algorithme peut prendre parfois des actions suboptimales afin de permettre à l'agent d'explorer l'espace d'actions, ce qui peut entraîner des récompenses moins importantes.

À la fin du processus d'entraînement, l'agent a engrangé 200 000 trajectoires et récompenses. Il va alors examiner toutes les récompenses pour déterminer quelle trajectoire a généré le montant le plus élevé et choisir l'allocation d'actifs de cette trajectoire qui a maximisé cette récompense.

Suite à la définition du protocole de l'étude ainsi que du calibrage du modèle, la prochaine étape suivante consiste en l'interprétation des résultats. Le chapitre suivant vise à analyser les performances des quatre allocations optimales générées en comparaison avec l'allocation initiale *Fixed-Mix* au travers des différentes métriques définies en partie I.7. L'analyse se concentre en premier lieu sur la répartition entre les différents actifs de chaque stratégie. Puis dans un second temps, il est étudié l'impact sur les différents indicateurs et sur les différents mécanismes du modèle épargne.

Ce chapitre permet de fournir une étude comparative entre les différentes allocations et de valider la robustesse de la méthode, mais aussi de pouvoir distinguer les avantages et les limites des stratégies basées sur le *reinforcement learning*.

Chapitre IV

Analyse et interprétation des résultats



L'objectif de ce chapitre est de présenter les différents résultats obtenus dans le cadre de cette étude. Pour rappel, quatre stratégies d'allocation ont été retenues au total : trois dans un univers d'entraînement risque neutre et une dans un univers d'entraînement risque réel. Par la suite, ces stratégies seront comparées à la stratégie actuelle *Fixed-Mix* dans un univers risque neutre, en utilisant les critères définis précédemment.

IV.1 Répartition des actifs dans les stratégies d'allocation

La sélection des portefeuilles optimaux consiste à sélectionner les portefeuilles optimisant les indicateurs suivants :

— **Stratégie $SCR_{marché}$:**

- Entraînement risque neutre
- Métrique : SCR de marché (à minimiser)
- Objectif : Réduire le SCR de marché pour diminuer le besoin en capital réglementaire.

— **Stratégie PVFP :**

- Entraînement risque neutre
- Métrique : PVFP (à maximiser)
- Objectif : Augmenter la PVFP pour améliorer la rentabilité des contrats.

— **Stratégie PVFP/ $SCR_{marché}$:**

- Entraînement risque neutre
- Métriques : SCR de marché (à minimiser) et PVFP (à maximiser)
- Objectif : Trouver une optimisation permettant la réduction du SCR de marché tout en augmentant le plus possible la PVFP.

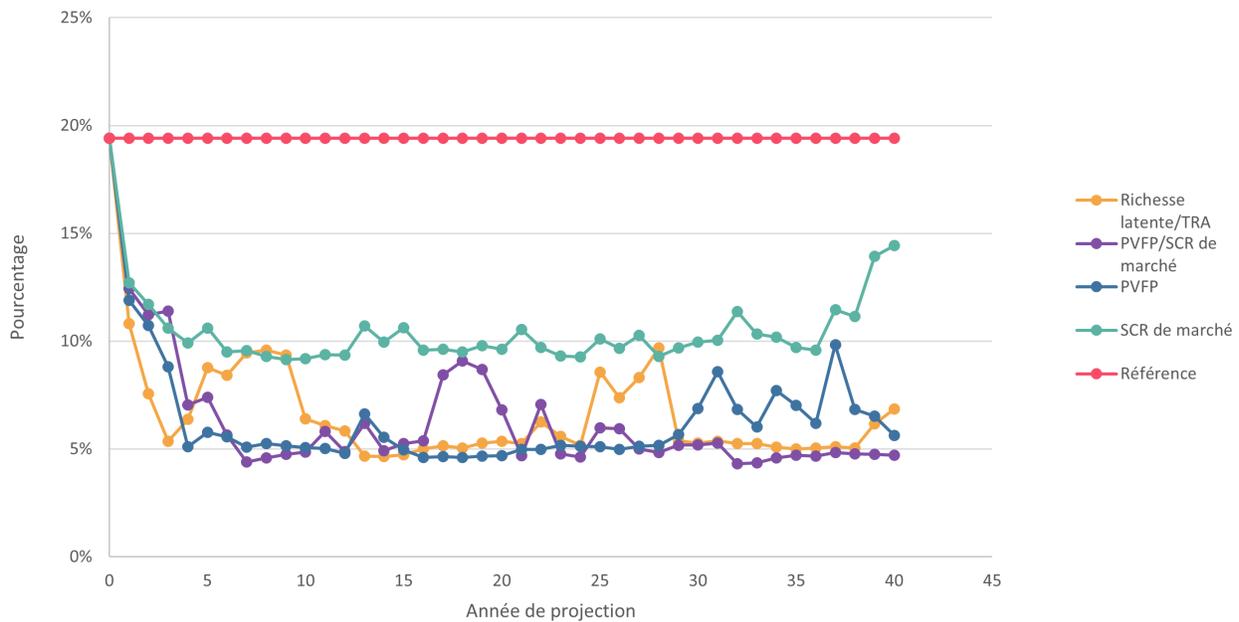
— **Stratégie Richesse latente/TRA :**

- Entraînement risque réel
- Métriques : TRA (à maximiser) et Richesse Latente (à maximiser)
- Objectif : Maximiser le TRA pour améliorer le rendement ajusté au risque et augmenter la richesse latente pour renforcer la santé financière à long terme. Cette stratégie permet ainsi d'observer le comportement de l'algorithme dans un univers autre que le risque neutre.

Pour rappel, le cadre de cette étude s'inscrit dans le contexte de Solvabilité II. Dans cette étude, les coûts de transaction ne sont pas pris en compte, ce qui simplifie l'analyse des stratégies d'allocation d'actifs, mais peut également produire des résultats peu exploitables dans le monde réel. Cependant, cela permet de faire une étude comportementale de l'algorithme dans un contexte normatif.

Par la suite, l'analyse vise à étudier en détail les résultats des stratégies d'allocations par classes d'actifs. Après avoir été entraîné dans des univers neutre ou réel, les agents fournissent une stratégie d'allocation d'actifs en univers risque neutre afin de pouvoir les comparer.

Analyse de la poche actions par stratégie



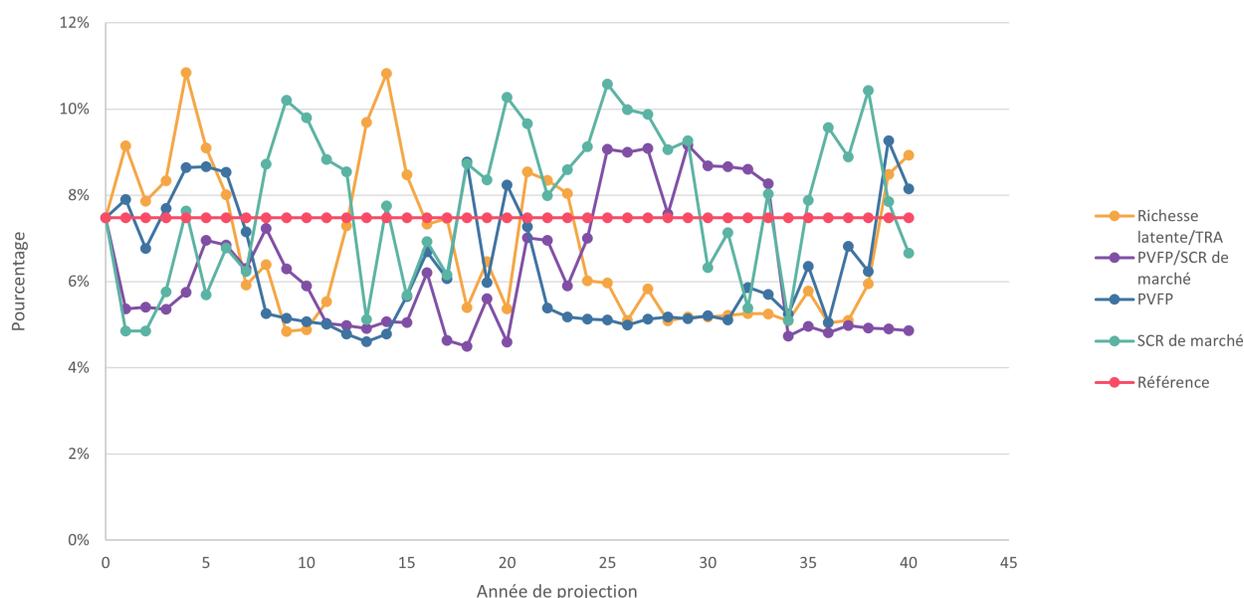
Stratégie	Moyenne des actions	Variance des actions	Écart-Type des actions
Référence	19.40%	0.00%	0.00%
PVFP/SCR _{marché}	6.31%	8.47%	2.91%
PVFP	6.40%	7.25%	2.69%
SCR _{marché}	10.48%	3.45%	1.86%
Richesse latente/TRA	6.69%	7.03%	2.65%

FIGURE IV.1 – Pourcentages et statistiques de la part des actions au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs

La figure ainsi que le tableau IV.1 fournissent un aperçu comportemental de chacune des stratégies. Une première analyse permet de révéler une différence notable : les stratégies DDPG tendent à diminuer significativement la part des actions dans le portefeuille par rapport à la stratégie de référence. Il apparaît que, pour les actions, les contraintes sur la classe d'actif sont respectées, avec une variation d'une année à l'autre de la part des actions dans le portefeuille bien comprise entre $\pm 8\%$. Pour la contrainte de dépassement des limites, il est observé que l'algorithme, notamment pour la stratégie PVFP/SCR_{marché}, transgresse la borne minimale fixée à 5%. Ceci est en partie dû au caractère non absolu de la contrainte comme défini dans le tableau III.2. En effet, cette situation est directement liée

à la fonction de récompense utilisée dans l'algorithme DDPG. Si la pénalité associée au non-respect de la contrainte minimale d'actions (5 %) n'est pas suffisamment prononcée dans la fonction de récompense, l'agent peut privilégier d'autres aspects jugés plus importants selon les critères de récompense et ainsi opter pour une part d'actions dans le portefeuille en deçà du seuil fixé. La stratégie optimisant le SCR de marché se distingue particulièrement par une moyenne d'allocation en actions plus élevée par rapport aux autres stratégies avec une moyenne de 10 %.

Analyse de la poche immobilière par stratégie



Stratégie	Moyenne de l'immobilier	Variance de l'immobilier	Écart-Type de l'immobilier
Référence	7.48%	0.00%	0.00%
PVFP/SCR _{marché}	6.31%	2.31%	1.52%
PVFP	6.25%	1.90%	1.38%
SCR _{marché}	7.85%	2.97%	1.73%
Richesse latente/TRA	6.83%	3.01%	1.74%

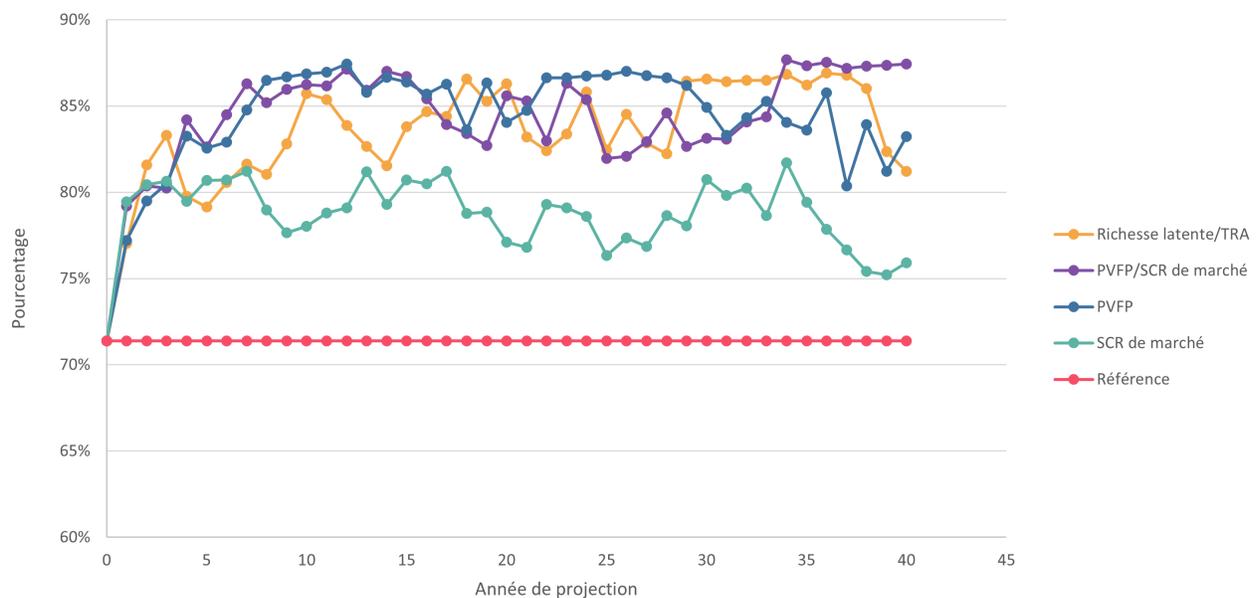
FIGURE IV.2 – Pourcentages et statistiques de la part de l'immobilier au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs

La figure et le tableau IV.2 illustrent la part de l'immobilier dans les différentes stratégies. De manière similaire aux actions, les stratégies respectent également les contraintes de variation annuelle pour l'immobilier, en restant dans la limite de $\pm 2\%$, ainsi que les bornes imposées, qui varient entre 5 % et 18 %. Dans le cas de l'immobilier, l'algorithme a tendance à osciller en termes de pourcentage autour de la stratégie de référence. Cependant, une exception notable est observée dans la stratégie PVFP, où la part d'immobilier est significativement inférieure entre les années de projections 8 et 38 par rapport

à la stratégie *Fixed-Mix*. Par la suite, une attention particulière sera accordée à l'explication de cette différence lors de l'interprétation des métriques. De plus, en admettant l'hypothèse de pouvoir acheter ou de vendre jusqu'à 2 % des parts immobilières à chaque pas de temps, il est crucial de surveiller les variations en valeur de marché de cette classe d'actif, surtout en tenant compte de l'illiquidité relative du marché de l'immobilier.

Analyse de la poche obligataire par stratégie

Le graphique IV.3 montre cette fois-ci que contrairement à la tendance observée pour les actions, l'algorithme DDPG a choisi d'augmenter la part des obligations dans le portefeuille pour la plupart des stratégies. Cependant, la stratégie $SCR_{marché}$ se distingue des autres. Elle maintient des proportions d'obligations plus proches de l'allocation de référence. Enfin, les contraintes associées à la classe d'actif sont respectées.



Stratégie	Moyenne des obligations	Variance des obligations	Écart-Type des obligations
Référence	71.39%	0.00%	0.00%
PVFP/SCR _{marché}	84.41%	9.13%	3.02%
PVFP	84.38%	9.86%	3.14%
SCR _{marché}	78.70%	4.21%	2.05%
Richesse latente /TRA	83.51%	9.71%	3.12%

FIGURE IV.3 – Pourcentages et statistiques de la part des obligations au sein du portefeuille en fonction des différentes stratégies d'allocation d'actifs

L'ensemble de ces analyses permet de vérifier leur conformité aux contraintes et vérifier leur cohérence

globale. Les différents résultats obtenus démontrent également le respect de la contrainte de maintenir 3 % en cash dans le portefeuille.

Une première conjecture concernant la stratégie optimisant la PVFP est que le modèle privilégie une allocation moindre dans les actions. En effet, celles-ci servent le taux sans risque mais présentent une volatilité élevée qui les rend moins avantageuses en risque neutre. Pour y remédier, le modèle investit donc dans des obligations plus rémunératrices dans un contexte de taux élevés afin d'augmenter la PVFP.

D'autre part, le modèle optimisant le $SCR_{marché}$ semble réduire sa part d'actions par rapport à la stratégie de référence, mais dans une moindre mesure que la stratégie PVFP. Cette approche modérée vise à limiter la production excessive de PVFP et à maintenir un delta de PVFP moins important lors du choc action, contribuant ainsi à la réduction du $SCR_{marché}$ par le biais du SCR_{action} .

La stratégie PVFP/ $SCR_{marché}$ adopte une approche similaire à celle de la stratégie PVFP afin de maximiser sa propre PVFP. Pour réduire le $SCR_{marché}$, cette stratégie capitalise sur le renouvellement et l'augmentation de la part des obligations dans le portefeuille. En procédant ainsi, elle tire parti de taux d'intérêt plus élevés, ce qui contribue à une meilleure immunisation contre le SCR_{spread} . L'intégration de nouvelles obligations, offrant des taux plus avantageux, permet de compenser les effets d'un choc de spread sur la valeur initiale du portefeuille, atténuant ainsi l'impact sur le delta de la PVFP et contribuant efficacement à la diminution du $SCR_{marché}$.

La stratégie risque réel TRA/Richesse latente adopte une approche similaire aux stratégies visant à augmenter la PVFP, en cherchant à maximiser les produits financiers à travers une allocation significative en obligations. Cette démarche permet d'avoir TRA plus élevé.

La suite de cette étude se tourne désormais vers une explication approfondie de chacune de ces stratégies à l'aide de différentes métriques. La stratégie en univers risque réel sera traitée séparément des autres stratégies définies dans un univers risque neutre de par la disjonction d'univers rendant leur comparaison incohérente. En effet, en raison de son univers d'entraînement, une comparaison directe avec les autres stratégies risque neutre n'est pas pertinente.

IV.2 Impacts des résultats sur les différentes métriques en entraînement risque neutre

L'objectif de cette section est de présenter les performances des stratégies d'allocation d'actifs entraînées en risque neutre au regard des différentes métriques.

IV.2.1 Stratégie $SCR_{marché}$

La première stratégie étudiée est celle ayant pour objectif d'optimiser la baisse du $SCR_{marché}$ avec un entraînement dans un univers risque neutre.

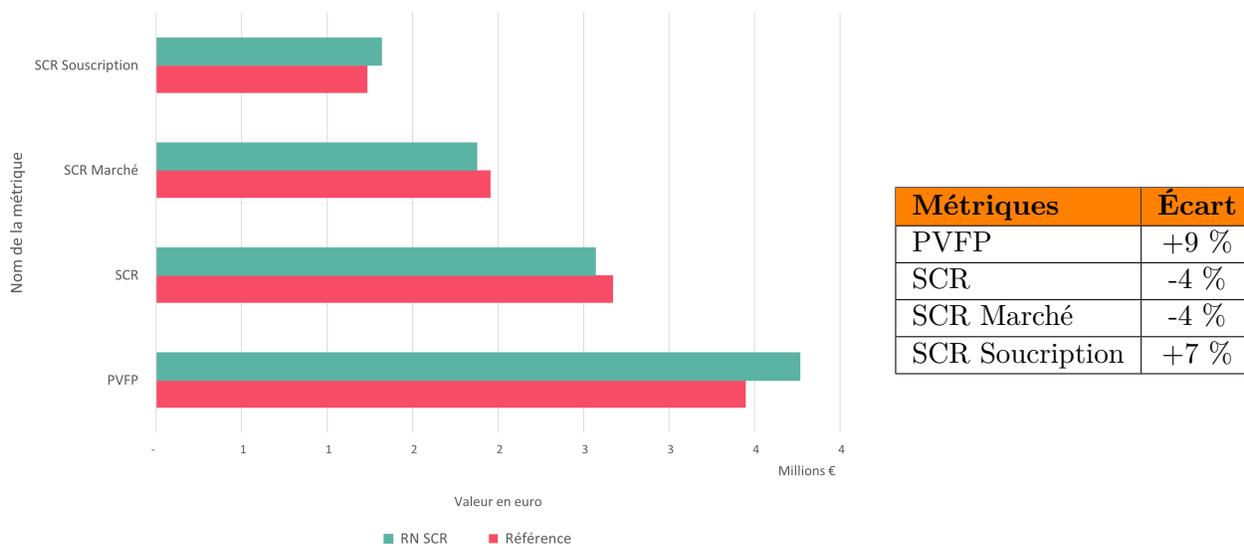
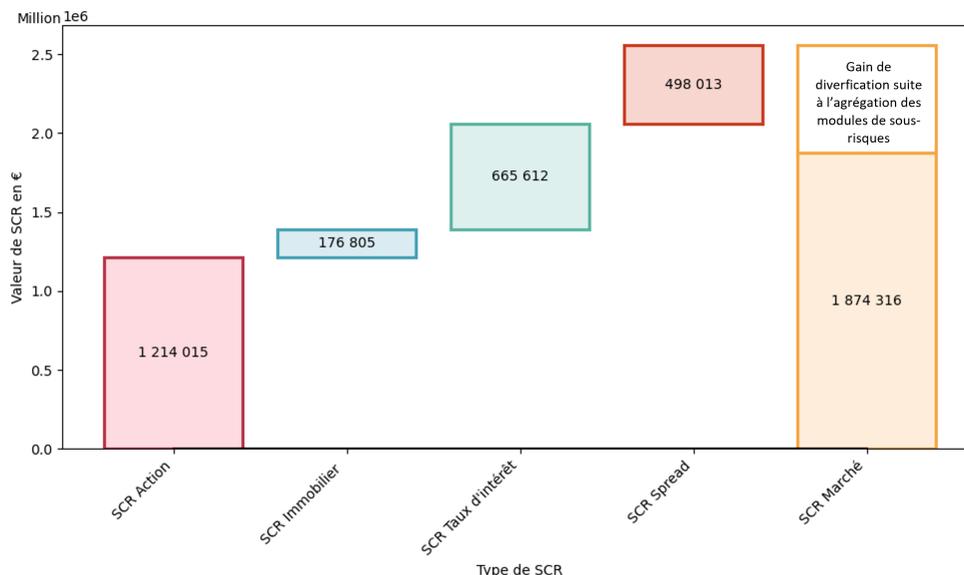


FIGURE IV.4 – Comparaison des métriques avec écart en % par rapport au scénario de référence

Les figures IV.4 récapitulent les différents résultats obtenus à l'aide de cette stratégie. Pour rappel, cette stratégie affiche une moyenne de part d'obligations de 78 %, ce qui est nettement supérieur à la moyenne de la stratégie de référence. Cette répartition est cohérente avec les objectifs de cette stratégie car cela entraîne en contrepartie une diminution de la part des actions dans le portefeuille, ce qui contribue ainsi à diminuer le $SCR_{marché}$, dans ce cas de -4 %.



Composante du SCR	SCR action	SCR immobilier	SCR taux	SCR spread
Écart en € et en %	-41 653 (-3,3 %)	-21 006 (-10,6 %)	58 439 (9,6 %)	-57 857 (-10,4 %)

FIGURE IV.5 – Décomposition du SCR de marché et écart en % pour la stratégie d'optimisation $SCR_{marché}$ par rapport à la stratégie de référence

Le graphique IV.5 offre une perspective plus granulaire, révélant que la réduction du $SCR_{marché}$ résulte principalement d'une baisse significative du SCR_{action} et du SCR_{spread} . La baisse du SCR_{action} , comme expliqué précédemment, est due à la diminution de la part des actions dans le portefeuille, tandis que la diminution du SCR_{spread} est liée à l'augmentation de la part d'obligations en portefeuille comme le montre la figure IV.7. En effet, les anciennes obligations, générant des coupons négatifs, sont vendues lorsqu'elles atteignent leur maturité ou alors vendues en moins-values latentes (MVL). Elles sont remplacées par de nouvelles obligations en plus grande quantité, bénéficiant également de taux plus avantageux. Elles sont gardées en portefeuille jusqu'à maturité, voir revendues en plus-values latentes (PVL). Cela permet, par rapport à la stratégie de référence, de dégager un résultat financier plus important avec ou sans choc de spread (cf. IV.6). De plus, ce mécanisme permet de réduire la sensibilité d'un choc de spread car l'assiette obligataire impactée par le choc est moins importante. Ce qui résulte d'un delta de PVFP moins conséquent et, entraîne une diminution du SCR_{spread} .

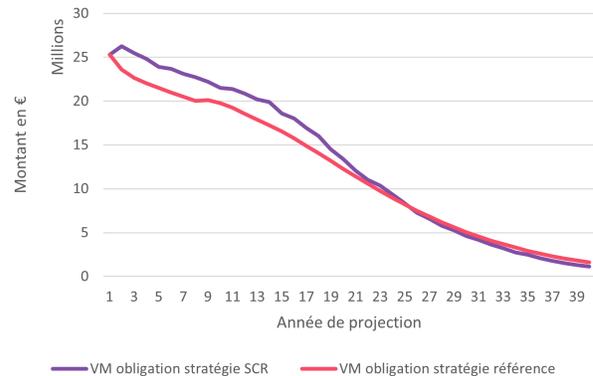
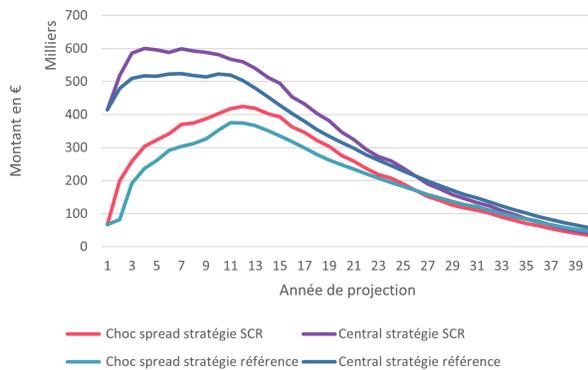


FIGURE IV.6 – Produits financiers en scénario central et choc de spread pour les deux stratégies

FIGURE IV.7 – Part en valeur de marché des obligations pour les deux stratégies

Cependant, l'achat supplémentaire d'obligations expose le portefeuille à un risque de hausse des taux. Le choc entraîne une augmentation du taux demandé par les assurés (TME) comme le montre la figure ci-dessous IV.8. Les obligations plus anciennes, déjà présentes dans le portefeuille, subissent alors des MVL. L'achat de nouvelles obligations à des taux plus élevés ne parvient pas à compenser cette différence, augmentant la probabilité de rachat par rapport au scénario de référence, engendrant une augmentation du $SCR_{\text{taux}}^{\text{up}}$ par effet volume.

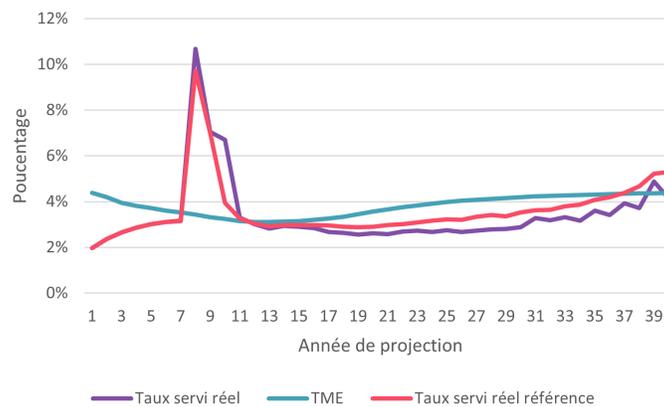


FIGURE IV.8 – TME et Taux servi réel pour la stratégie d'optimisation et de référence en scénario choc de taux up

Le SCR global connaît également une baisse de 4 %. Cette diminution est principalement due à la consolidation du $SCR_{\text{marché}}$. En revanche, le $SCR_{\text{souscription}}$ se dégrade, notamment à cause d'une forte hausse du $SCR_{\text{rachat}}^{\text{massif}}$, qui atteint 1 208 k€ contre 1 126 k€ dans le scénario de référence, soit une augmentation de 7 %. L'augmentation du $SCR_{\text{rachat}}^{\text{massif}}$ est notamment liée aux *management rules* adoptées par le modèle. En effet, durant les premières années de projection, le modèle réalise toutes les

plus-values latentes, générant ainsi un flux important de produits financiers. Cette stratégie augmente immédiatement le TRA en début de période, permettant d'alimenter abondamment la PPB.

La figure IV.9 met en évidence une augmentation significative du montant de la PPB durant les premières années, résultant d'une gestion moins lissée par le modèle. Cette hausse initiale de la PPB aide à limiter les rachats en début de projection, grâce à un taux servi aux assurés plus élevé, comme le montre la figure IV.10. Après 8 ans, une augmentation notable du taux servi est observée, répondant aux exigences réglementaires de redistribution des bénéfices.

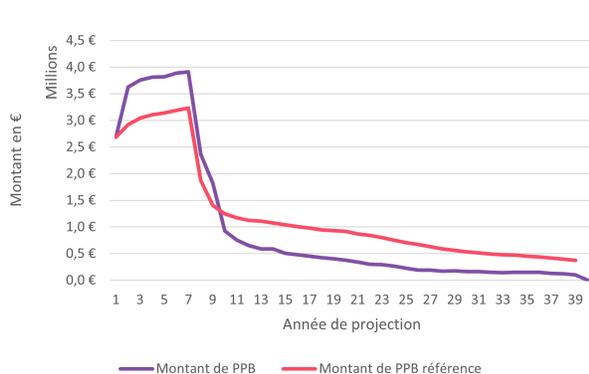


FIGURE IV.9 – Montant de PPB pour la stratégie d'optimisation et de référence

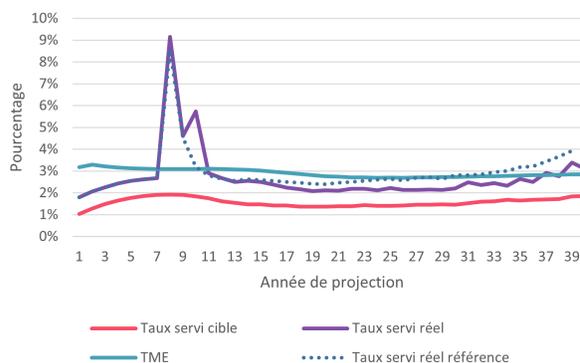


FIGURE IV.10 – TME, Taux servi et réel pour la stratégie d'optimisation et de référence en scénario central

Une gestion qui se concentre sur la vente massive de PVL des actions en début de période peut s'avérer problématique sur le long terme. En effet, cela implique après la redistribution de la PB réglementaire une réserve en PPB plus faible pouvant entraîner une baisse du taux servi, ce qui déclenche une hausse des rachats. La figure IV.10 illustre le fait qu'après 8 ans le taux servi réel est inférieur aux taux TME ce qui implique une augmentation des rachats conjoncturels. C'est pour cette raison que le $SCR_{rachats}$ se dégrade.

Malgré le fait que l'objectif de ce modèle ne soit pas d'améliorer la PVFP, celle-ci est positivement impactée, s'élevant dans cette stratégie à 3 765 k€, supérieur au montant de 3 449 k€ du scénario de référence. L'augmentation observée est encore directement liée à la réalisation massive des PVL sur les actions. Cette stratégie, qui a pour effet d'accroître le provisionnement pour la PPB et d'atténuer le taux servi moyen notamment à cause de la contrainte de la PB, permet ainsi d'augmenter in fine le résultat financier et, par conséquent, la PVFP.

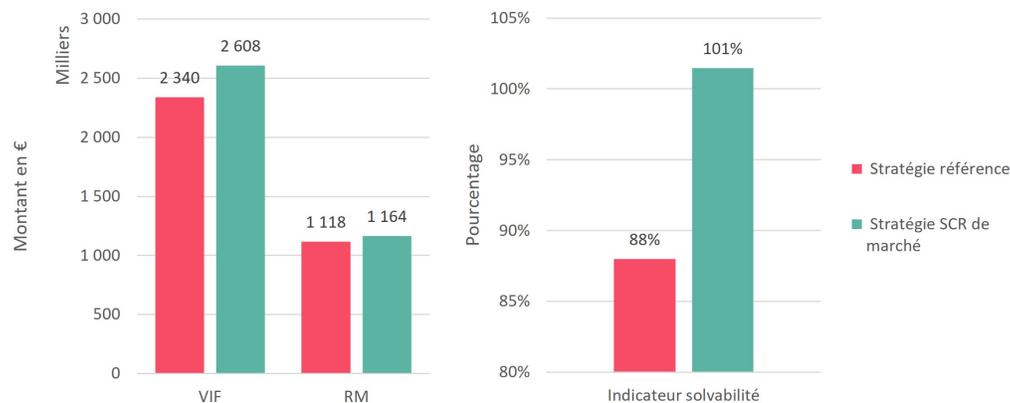


FIGURE IV.11 – VIF, RM ainsi que l'indicateur de solvabilité de l'allocation de référence et celle optimisant le SCR de marché

Une hausse de la VIF est également observée, due à une augmentation plus significative de la PVFP, partiellement compensée par une augmentation de la marge de risque liée à l'augmentation du $SCR_{souscription}$. Cependant, la consolidation du SCR absorbe une partie des changements en matière de variations de la VIF mais permet toutefois d'obtenir une amélioration nette de l'indicateur de solvabilité.

Après avoir détaillé les nuances de la stratégie axée sur le $SCR_{marché}$, la prochaine partie est dédiée à la stratégie optimisant de la PVFP.

IV.2.2 Stratégie PVFP

La stratégie d'optimisation de la PVFP possède une part encore plus importante d'obligations de l'ordre de 84 % par rapport à l'allocation optimisant le $SCR_{marché}$ avec 78 % d'obligations tandis qu'elle diminue fortement la part des actions par rapport à la stratégie de référence. Pour chaque classe d'actifs au sein de cette stratégie l'algorithme démontre une tendance à sélectionner des allocations se situant à proximité des limites extrêmes définies. L'orientation de l'algorithme vers les bornes minimales ou maximales pour chaque classe d'actifs pourrait soulever des questions quant au phénomène de surapprentissage. En effet, en s'ajustant finement aux contraintes spécifiques du jeu de données d'entraînement, l'algorithme pourrait potentiellement perdre en généralisation.

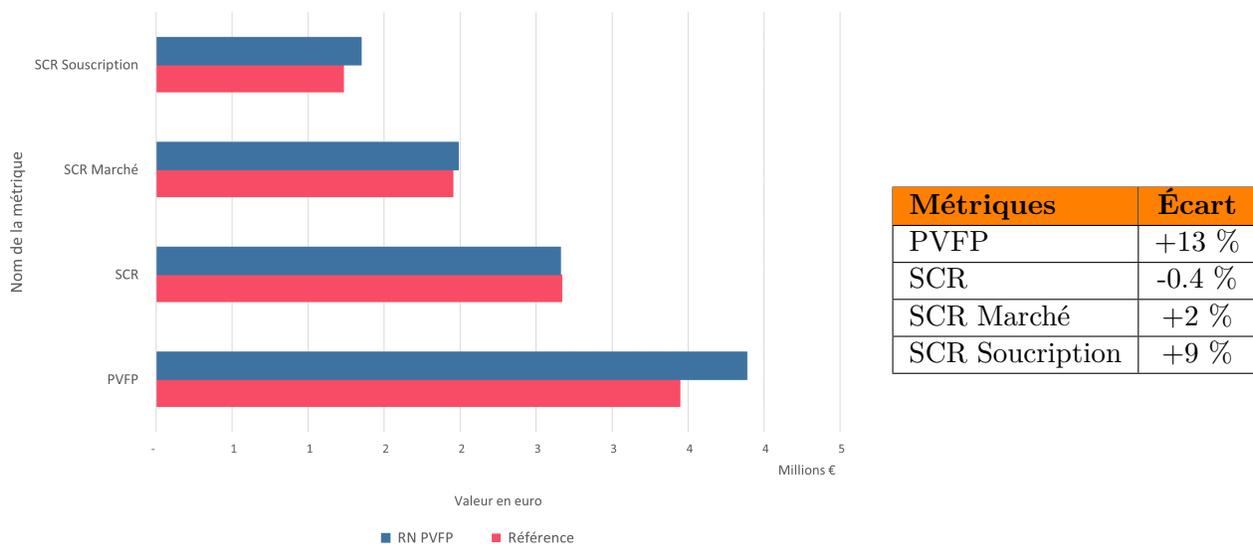
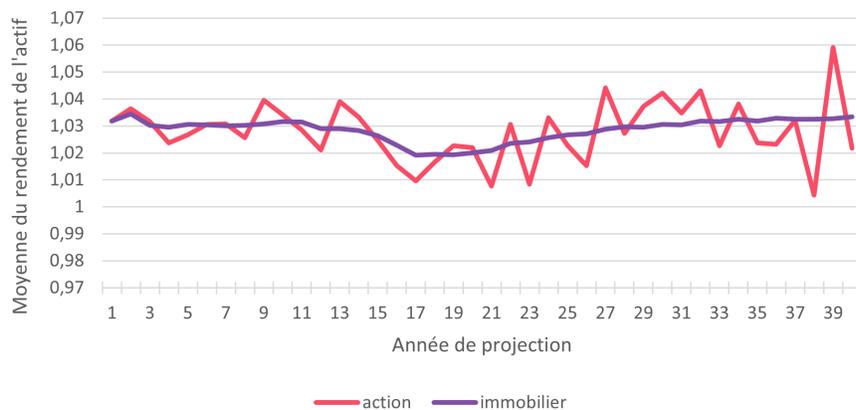


FIGURE IV.12 – Comparaison des métriques avec écart en % par rapport au scénario de référence

Les figures IV.12 présentent un résumé des résultats obtenus grâce à notre stratégie d’optimisation de la PVFP. Cette composition modifiée du portefeuille, en mettant un fort accent sur les obligations, a conduit à une augmentation notable de la PVFP de +13 % battant ainsi la performance de l’allocation optimisant le $SCR_{marché}$.

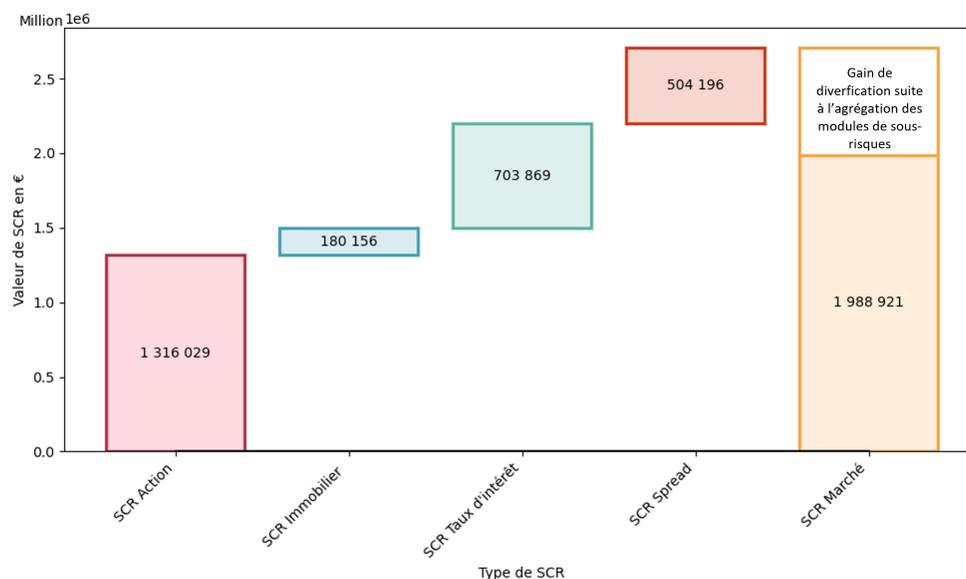
Cette augmentation est notamment due à la capacité du modèle d’arbitrer entre la part d’immobilier et d’actions dans la stratégie. En effet, comme le montrent les courbes, la part de l’immobilier et des actions dans le portefeuille s’entrecroisent lors des projections. Ce phénomène est lié au fait que le log-rendement en figure IV.13 de ces deux actifs est très proche, ce qui induit que le modèle change sa préférence entre l’immobilier et les actions.



Actif	Moyenne
Action	1,027
Immobilier	1,028

FIGURE IV.13 – Log-rendement des actions et de l’immobilier des projections risque neutre

L’amélioration de la PVFP dans cette stratégie d’optimisation est corrélée à trois facteurs : une augmentation de la part d’obligations permettant de profiter de taux attractifs et de PVL importantes, une vente brutale en début d’année de projection des actions comme le montre la figure IV.1 et enfin un arbitrage fin entre l’immobilier et les actions entraînant là aussi une augmentation des plus-values latentes et, in fine, de la PVFP.



Composante du SCR	SCR action	SCR immobilier	SCR taux	SCR spread
Écart en € et en %	60 361 (4,8 %)	-17 655 (-8,9 %)	96 696 (15,9 %)	-51 674 (-9,3 %)

FIGURE IV.14 – Décomposition du SCR de marché et écart en % pour la stratégie d’optimisation PVFP par rapport à la stratégie de référence

Le $SCR_{marché}$, comme le montre la figure IV.14, se détériore. Malgré une réduction de la part des actions dans la stratégie d'optimisation de la PVFP, cette stratégie se retrouve paradoxalement plus exposée au SCR_{action} que celle optimisant le $SCR_{marché}$. Cette situation découle principalement de la manière dont la stratégie PVFP cherche à accroître la PVFP, en s'appuyant de manière significative sur l'augmentation des résultats financiers grâce à une vente massive des PVL actions en début de projection. En revanche, la stratégie optimisant le SCR de marché, maintiens une part stable d'actions à environ 10 % et ne liquide pas toutes ses PVL. Ainsi, la stratégie optimisant la PVFP est exposée à un delta plus important lors d'un choc action impliquant un SCR_{action} plus important.

La stratégie PVFP suit le même mécanisme que la stratégie $SCR_{marché}$. Cependant, elle possède une part encore plus importante d'obligations, ce qui l'expose à une variation de taux haussier du fait d'un effet volume plus important impliquant un SCR_{taux}^{UP} qui se dégrade.

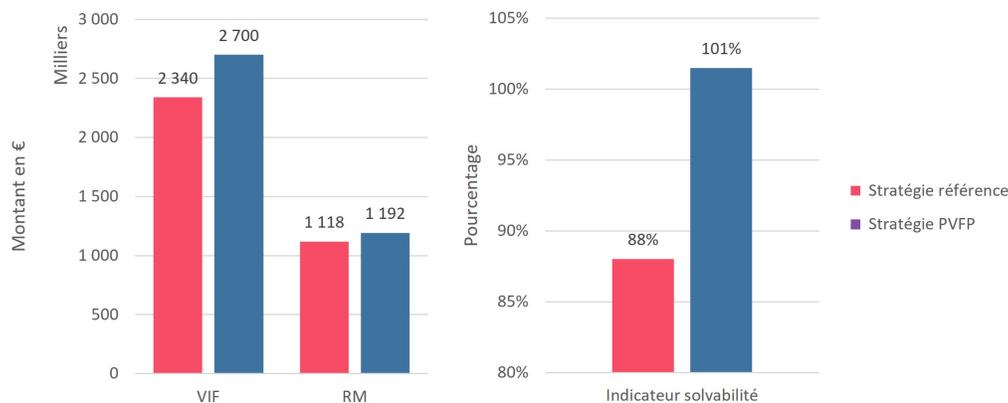


FIGURE IV.15 – VIF, RM ainsi que l'indicateur de solvabilité de l'allocation de référence et celle optimisant la PVFP

Malgré une dégradation du $SCR_{marché}$ (cf. IV.12), l'indicateur de solvabilité reste stable par rapport au résultat précédent optimisant le $SCR_{marché}$, par le fait que la PVFP est consolidée de manière à contrebalancer l'effet négatif d'une diminution du $SCR_{marché}$.

L'explication de la stratégie optimisant de manière distincte la PVFP et celle optimisant le $SCR_{marché}$ conduit à se pencher vers un modèle hybride optimisant la PVFP et le $SCR_{marché}$. Cette orientation vers un modèle combiné s'inscrit dans une perspective de pilotage stratégique, où l'objectif est de parvenir à un équilibre optimal entre la performance financière et la gestion réglementaire.

IV.2.3 Stratégie $SCR_{marché}$ et PVFP

L'étude se poursuit avec la stratégie qui vise à optimiser simultanément la PVFP et le $SCR_{marché}$. Cette approche permet d'observer comment le modèle interagit avec son environnement afin d'optimiser deux métriques.

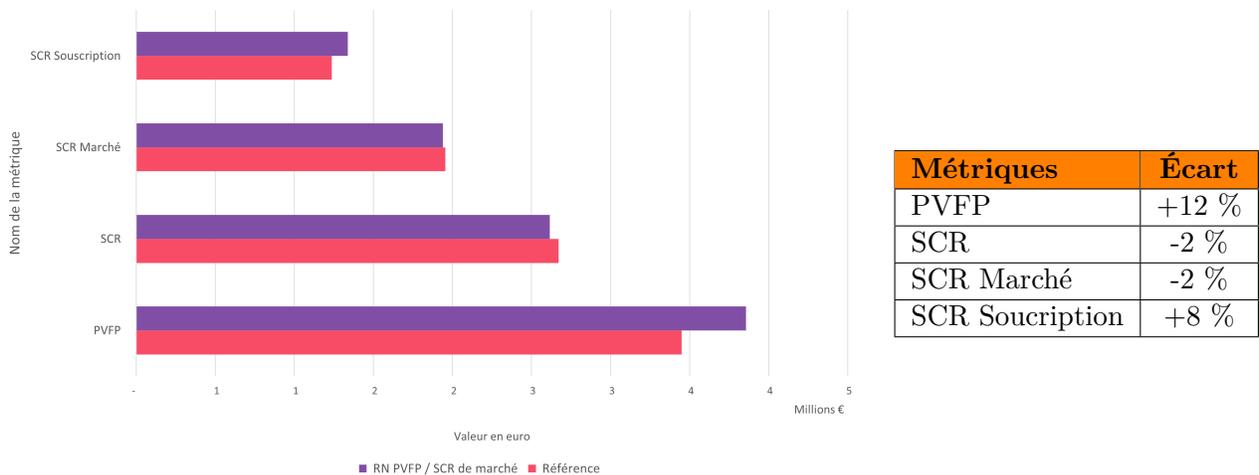
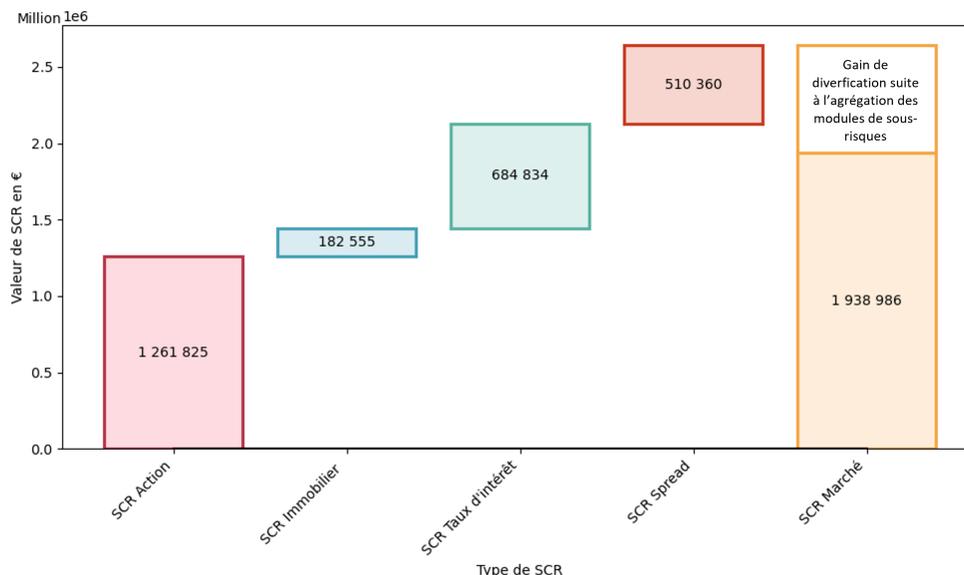


FIGURE IV.16 – Comparaison des métriques avec écart en % par rapport au scénario de référence

La figure IV.16 expose les résultats de la stratégie. Une première remarque est de constater que les deux objectifs de la stratégie d'augmenter la PVFP et de diminuer le $SCR_{marché}$ sont atteints avec une augmentation de +12 % pour la PVFP et une diminution de -2 % du $SCR_{marché}$. De plus, il est intéressant de souligner que la performance isolée de la PVFP (3 857 k€) n'a pas été aussi significative que celle observée dans le modèle exclusivement axé sur l'optimisation de la PVFP (3 885 k€). De même, la réduction du $SCR_{marché}$ (1 938 k€) n'a pas été aussi prononcée que dans le modèle dédié uniquement à l'optimisation du $SCR_{marché}$ (1 874 k€). Cette moindre performance est attribuable au fait que le poids de la récompense attribuée pour l'amélioration de chaque métrique est identique, contraignant ainsi le modèle à effectuer des compromis entre les objectifs.

Il apparaît que le modèle adopte dans une optique d'améliorer la PVFP et le $SCR_{marché}$ un comportement similaire en termes de stratégie d'allocation d'actifs que la stratégie optimisant la PVFP. Ainsi les parts d'actifs de ces deux stratégies sont très proches.

La PVFP de cette stratégie est améliorée grâce à un mécanisme similaire à celui de la stratégie PVFP, impliquant un arbitrage entre les actions et l'immobilier.



Composante du SCR	SCR action	SCR immobilier	SCR taux	SCR spread
Écart en € et en %	6 157 (0,5 %)	-15 256 (-7,7 %)	77 661 (12,7 %)	-45 510 (-8,2 %)

FIGURE IV.17 – Décomposition du SCR de marché et écart en % pour la stratégie d'optimisation PVFP/SCR_{marché} par rapport à la stratégie de référence

Le graphique IV.17 illustre la décomposition du SCR_{marché} et révèle que la réduction du SCR_{marché} est principalement due à une amélioration du SCR_{spread}. Cette amélioration est rendue possible par le renouvellement de la part obligataire du portefeuille avec de nouvelles obligations offrant des taux plus favorables. Par ailleurs, le SCR_{immobilier} connaît également une diminution, attribuable à la réduction moyenne de la part d'immobilier par rapport à la stratégie de référence. Bien que la part des actions reste similaire à celle observée dans la stratégie optimisant uniquement la PVFP, le SCR_{action} est moins impacté. Cette différence s'explique par une assiette de PVFP moins conséquente, rendant le portefeuille moins sensible aux chocs sur les actions.

La consolidation du SCR_{marché} ainsi qu'une dégradation moins importante du SCR_{souscription} 1 338 k€ contre 1 348 k€ pour la stratégie optimisant la PVFP, permet in fine, d'améliorer le SCR.

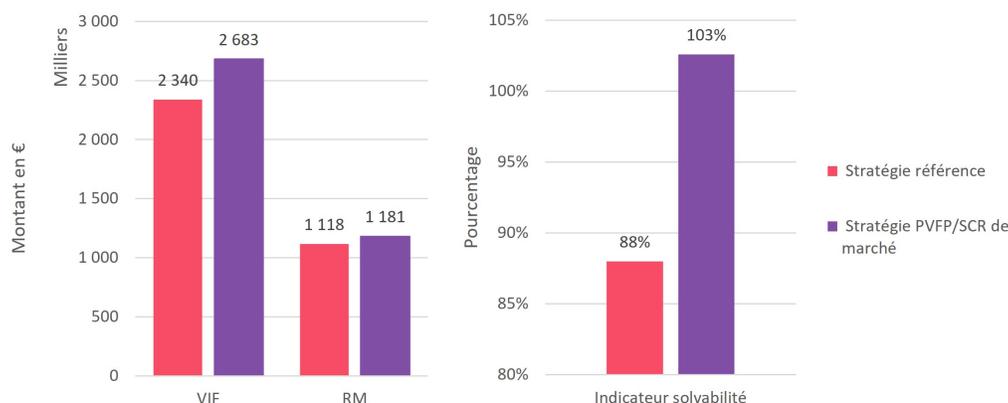


FIGURE IV.18 – VIF, RM ainsi que l'indicateur de solvabilité de l'allocation de référence et celle optimisant le SCR de marché et la PVFP

L'indicateur de solvabilité de cette stratégie se révèle être le plus important (103 %) parmi les trois approches d'optimisation, par rapport à la stratégie de référence. Cette performance est due à un double effet : d'une part, il s'agit de la seconde PVFP la plus élevée de toutes les stratégies, permettant d'avoir une VIF importante, d'autre part, contrairement à la stratégie purement axée sur la PVFP, l'approche qui optimise à la fois la PVFP et le $SCR_{marché}$ aboutit à un SCR plus élevé. Ces deux composantes permettent d'obtenir l'indicateur de solvabilité le plus solide.

Récapitulatif :

Stratégies / Métriques	Référence	RN PVFP/ $SCR_{marché}$	RN PVFP	RN $SCR_{marché}$
PVFP	3 449 110	3 857 009 ↑	3 885 879 ↑	3 765 291 ↑
SCR	2 672 117	2 614 645 ↓	2 660 763 ↓	2 569 768 ↓
SCR Marché	1 955 249	1 938 986 ↓	1 988 921 ↑	1 874 316 ↓

TABLE IV.1 – Résumé de l'efficacité des stratégies sur les différentes métriques par rapport au scénario de référence

Après avoir présenté l'ensemble des résultats par stratégie, le récapitulatif dans le tableau ci-dessus IV.1 met en évidence la polyvalence des stratégies $SCR_{marché}$ et PVFP/ $SCR_{marché}$, comme le suggèrent les variations observées sur les différentes métriques. Dans la suite de cette étude, la stratégie PVFP/ $SCR_{marché}$ sera employée pour conduire diverses analyses de sensibilité économique et de modèle, afin d'évaluer sa robustesse face à des scénarios variés et de continuer l'analyse comportementale du *reinforcement learning* sur ses performances.

L'étude se poursuit avec l'analyse des résultats de la stratégie après un entraînement dans un contexte de risque réel, offrant un aperçu du comportement face à un autre univers de projection.

IV.3 Impacts des résultats sur les différentes métriques en entraînement risque réel

La stratégie en risque réel a pour but d'optimiser le TRA et la richesse latente. Elle a été entraînée dans un environnement réel puis, comme les trois stratégies précédentes, testée dans un contexte neutre pour permettre une comparaison avec les autres stratégies. Cette approche a pour objectif d'examiner le comportement du *reinforcement learning* face à un autre univers de projection. Ainsi, les résultats de cette partie sont présentés en risque neutre.

Une première observation est que la stratégie adopte des proportions d'allocation similaires à celles observées dans les stratégies qui se concentrent sur l'optimisation de la PVFP et le couple $SCR_{marché}/PVFP$. Une des conséquences attendues d'une telle allocation est un résultat financier amélioré, qui permet d'accentuer le TRA.

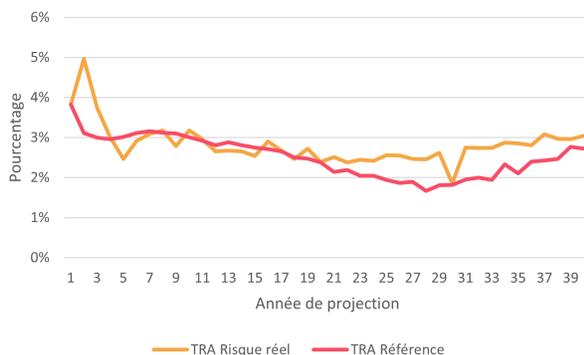


FIGURE IV.19 – TRA pour la stratégie d'optimisation

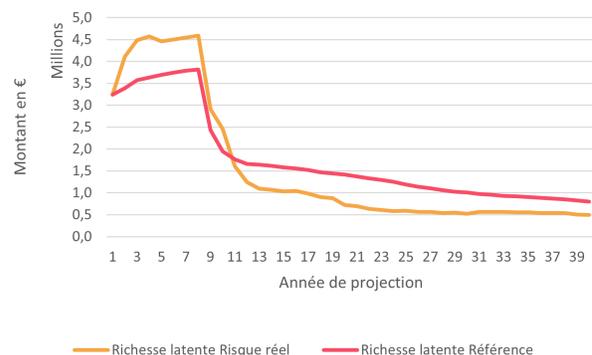


FIGURE IV.20 – Richesse latente pour la stratégie d'optimisation

Stratégie	Référence	Stratégie TRA/Richesse latente
TRA	2,25%	2,82%
Richesse latente	1 742 075 €	1 543 429 €

TABLE IV.2 – Comparaison entre la stratégie de référence et la stratégie TRA/Richesse latente

La figure ci-dessus IV.19 illustre que le modèle réussit à générer un TRA plus élevé grâce à l'optimisation de la stratégie d'allocation d'actifs. Ceci est dû au fait que le modèle mise sur une maximisation

du produit financier pour obtenir un TRA supérieur. Cette maximisation résulte d'une vente massive des PVL actions. Cependant, la richesse latente se détériore, affichant une baisse de -11 %. Cette situation est principalement due à deux phénomènes :

- L'univers d'entraînement étant à risque réel, il existe un risque de perte d'informations lors du passage à un univers risque neutre, ce qui peut entraîner une dégradation de la qualité du modèle.
- Le caractère antinomique des deux métriques peut contraindre le modèle à réaliser des arbitrages, en favorisant l'une au détriment de l'autre pour maximiser sa récompense.

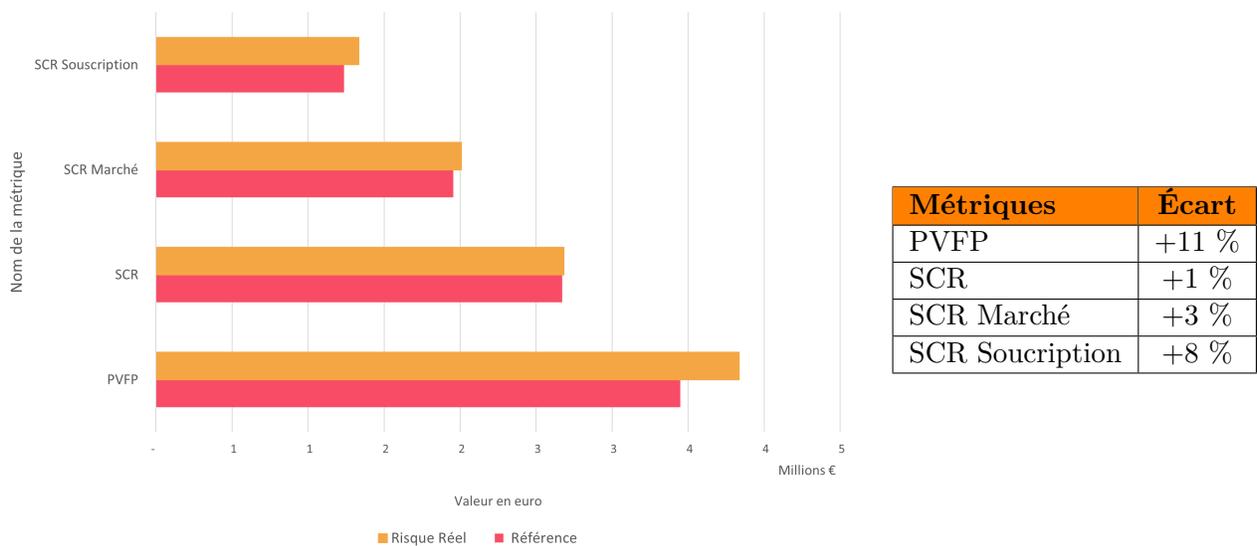


FIGURE IV.21 – Comparaison des métriques avec écart en % par rapport au scénario de référence

La figure ci-dessus IV.21 illustre que l'accent mis sur le TRA conduit à une amélioration significative du résultat financier, se traduisant par une augmentation de la PVFP de +11 %. Enfin, il apparaît au vu des résultats sur les différentes métriques Solvabilité II que cette méthode d'allocation d'actifs s'avère peu robuste dans ce cadre réglementaire.

Après avoir détaillé les différentes stratégies, l'attention se porte désormais sur les études de sensibilité et de robustesse du modèle. Ces analyses seront classées en deux catégories : les sensibilités liées au modèle lui-même et celles d'ordre économique.

IV.4 Analyse de sensibilité du modèle

L'objectif principal de cette section est d'étudier le comportement du modèle en ajustant divers hyperparamètres, afin d'étudier leur influence sur la performance et la stabilité du système. Trois analyses de sensibilité distinctes seront menées : la première se concentrera sur l'impact du nombre d'épisodes d'entraînement, la deuxième évaluera les effets de modifications apportées à la structure de récompense, et la troisième explorera les conséquences d'une optimisation sur le $SCR_{souscription}$ à la place du $SCR_{marché}$.

IV.4.1 Sensibilité sur le nombre d'épisodes d'entraînement

Le but de cette partie est d'observer l'impact du nombre d'épisodes sur la phase d'entraînement du modèle. Dans une optique d'industrialisation du modèle et face à des contraintes temporelles, il est essentiel de considérer le temps de calcul, ainsi que la stabilité et la convergence du modèle. La détermination du nombre optimal d'épisodes d'entraînement permet non seulement d'assurer une efficacité opérationnelle, mais également de garantir que le modèle atteint une performance satisfaisante sans sur-apprentissage ni sous-apprentissage. Pour cela, le modèle optimisant la PVFP/ $SCR_{marché}$ avec 200 épisodes d'entraînement sera considéré dans cette section comme le scénario de référence et sera comparé à trois autres modèles entraînés sur 50, 150 et 250 épisodes.

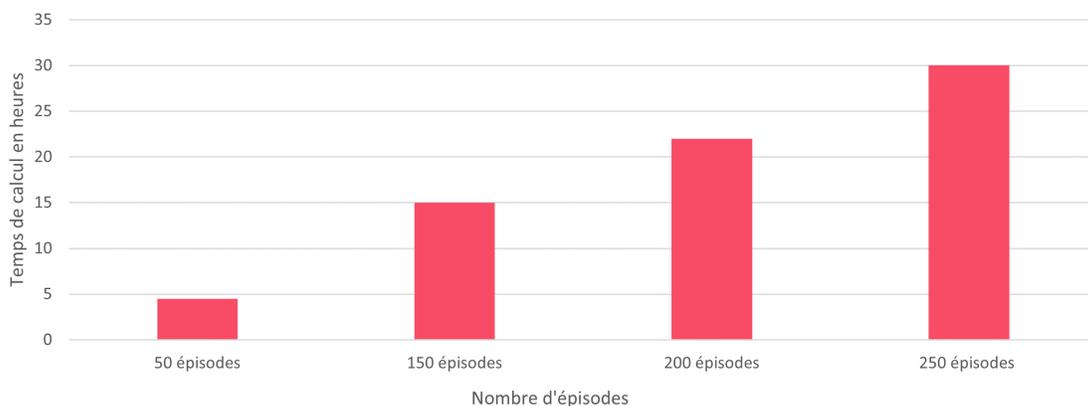


FIGURE IV.22 – Temps de calcul en fonction du nombre d'épisodes d'entraînement

Le graphique IV.22 illustre une relation non linéaire (quadratique) entre le temps de calcul et le nombre d'épisodes. Ce phénomène est notamment lié à la nature même du *reinforcement learning*, où l'accumulation de données d'expérience, due à la technique d'*experience replay*, augmente progressivement le temps de calcul. De plus, la convergence vers des politiques optimales rend les ajustements nécessaires plus subtils et exigeants en termes de calcul.

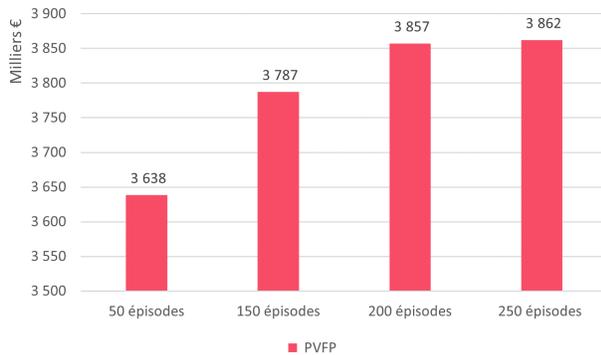


FIGURE IV.23 – PVFP en fonction du nombre d'épisodes d'entraînement

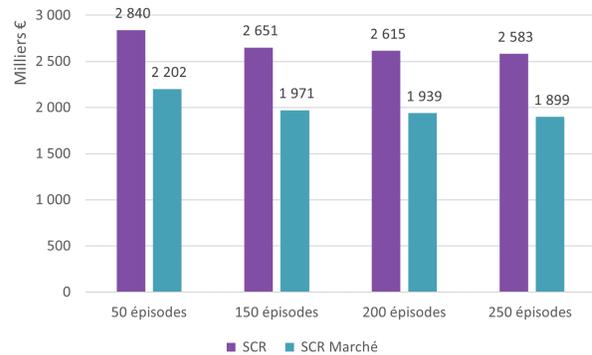


FIGURE IV.24 – SCR et SCR de marché en fonction du nombre d'épisodes d'entraînement

Les figures IV.23 et IV.24 présentent l'évolution des différentes métriques d'optimisation en fonction du nombre d'épisodes d'entraînement. Initialement, au début du processus d'entraînement, la PVFP, le SCR et le $SCR_{marché}$ sont dégradés, reflétant l'adaptation initiale du modèle à l'environnement et à ses contraintes. Cependant, à mesure que l'entraînement progresse, une amélioration graduelle de ces métriques est observée, signe que le modèle commence à identifier et à adopter des stratégies plus efficaces. Cette tendance à l'amélioration se poursuit jusqu'aux entraînements de 200 et 250 épisodes, où les métriques semblent se stabiliser.

Actif	50 ⇒ 150 épisodes	150 ⇒ 200 épisodes	200 ⇒ 250 épisodes
Obligations	0,22%	0,17%	0,03%
Actions	1,86%	1,65%	0,23%
Immobilier	0,67%	0,80%	0,02%

TABLE IV.3 – Écart relatif moyen d'un passage d'allocation à une autre pour chaque catégorie d'actifs

Les deux premières colonnes montrent des écarts relatifs moyens similaires illustrant ainsi le fait que le modèle recherche et trouve de nouvelles allocations optimales au cours de l'entraînement. Ensuite, sur le passage de 200 à 250 épisodes l'écart relatif diminue de manière significative. Cette réduction de l'écart relatif suggère que le modèle commence à se stabiliser. En effet, bien que le nombre d'épisodes continue d'augmenter et que le bruit sur le modèle soit toujours présent, les variations des pourcentages d'actifs obtenus sont moins marquées. Cela suggère que, même si l'algorithme parvient encore à identifier des allocations légèrement différentes, les améliorations en termes de métriques ciblées deviennent moins importantes.

Les différents résultats illustrés ci-dessus permettent, dans une perspective d'industrialisation, d'avoir une idée claire de l'impact du nombre d'épisodes sur l'entraînement du modèle. Ainsi, il semble optimal, notamment grâce à la stabilité et à la convergence du modèle, associé au temps de calcul, de

s'arrêter à 200 épisodes. Cette décision est basée sur le constat que les gains marginaux obtenus en augmentant le nombre d'épisodes sont limités, même s'il n'est pas exclu que le modèle puisse trouver des stratégies plus optimales avec un nombre d'épisodes supérieur. Cependant, cela se ferait au coût d'un temps de calcul plus important.

Après avoir étudié le comportement du modèle par rapport au nombre d'épisodes d'entraînement, la prochaine section discute des effets d'un autre hyperparamètre du modèle, qui est la structure de la récompense.

IV.4.2 Sensibilité sur la structure de récompense du modèle

Cette sensibilité a pour but d'étudier la performance du modèle en face à une modification de la structure de la récompense. Pour cela, la méthodologie suivante a été adoptée, elle consiste à introduire des poids aux différentes métriques dans la fonction de récompense, à l'aide des coefficients β_1 et β_2 modifiant ainsi la formule de la récompense ci-dessous IV.4.2. La stratégie optimisant la PVFP/SCR_{marché} est considérée comme celle de référence. Les coefficients modulent l'importance relative de chaque métrique dans la détermination de la récompense globale attribuée au modèle pour chaque décision prise au cours de l'entraînement. L'objectif est d'explorer comment le modèle priorise les métriques en fonction des poids attribués et d'évaluer l'impact de ces poids sur la performance globale du modèle.

$$\text{Récompense}_{\text{PVFP}} = \beta_1 \times m \times |\text{PVFP}(t, i) - \text{PVFP}(0, i)|, \quad \text{si } \text{PVFP}(t, i) > \max_{j < i} \text{PVFP}(t, j)$$

ou

$$\text{Récompense}_{\text{SCR}_{\text{marché}}} = \beta_2 \times m \times |\text{SCR}_{\text{marché}}(t, i) - \text{SCR}_{\text{marché}}(0, i)|, \quad \text{si } \text{SCR}_{\text{marché}}(t, i) < \min_{j < i} \text{SCR}_{\text{marché}}(t, j)$$

Pour étudier l'impact de la structure de récompense sur la performance du modèle, différentes combinaisons de poids ont été attribuées aux métriques PVFP et SCR_{marché}, comme illustré dans le tableau IV.4. La stratégie de "Référence" reprend la même structure de récompense qu'utiliser dans le chapitre IV.2.3, attribuant une importance égale à la PVFP et au SCR_{marché} avec des poids de 0.5. Le but de l'analyse est de choquer ces poids afin de favoriser une métrique par rapport à une autre. Ainsi, la combinaison qui "privilégie la PVFP" augmente le poids de la PVFP à 0.75, réduisant celui du SCR_{marché} à 0.25. Pour observer si le modèle favorise l'amélioration de la PVFP lorsque celle-ci est pondérée plus lourdement. La même méthodologie pour "privilégie le SCR_{marché}" est appliquée.

Combinaison	β_1 (PVFP)	β_2 (SCR _{marché})
Référence	0.5	0.5
Privilège PVFP	0.75	0.25
Privilège SCR _{marché}	0.25	0.75

TABLE IV.4 – Valeurs des coefficients pour différentes combinaisons dans la stratégie PVFP/SCR_{marché}

Le tableau ci-dessous IV.5 illustre les variations des métriques en fonction des différentes stratégies d'optimisation, au travers de la sensibilité axée sur la structure de récompense.

Métrique \ Stratégie	SCR	SCR de marché	PVFP
Référence	2 614 645	1 938 986	3 857 009
Privilège PVFP	2 612 564	1 932 831	3 853 621
Privilège SCR de marché	2 585 805	1 903 632	3 855 377

TABLE IV.5 – Résultats de la sensibilité sur la structure de récompense

En ce qui concerne la stratégie qui privilégie SCR_{marché}, une amélioration est observée, tout tout en diminuant de 1,8 % la PVFP par rapport à la PVFP de référence. Cela indique que le modèle, par le biais des poids, capte bien l'objectif de favoriser le SCR_{marché} par rapport à la PVFP. En contraste, la stratégie favorisant la PVFP affiche des performances moins bonnes que celles de la stratégie de référence, suggérant que l'accent mis sur la PVFP n'a pas induit de variations majeures dans les résultats du modèle par rapport à une approche plus équilibrée.

Ces résultats suggèrent une possible saturation dans l'optimisation de la PVFP. En effet, il semble que la PVFP est atteinte un plateau d'optimisation, rendant les améliorations difficiles sans ajustements majeurs dans l'allocation d'actifs ou la méthodologie. Parallèlement, la complexité intrinsèque du SCR_{marché}, comparativement à la PVFP, rend sa maîtrise moins directe, surtout lorsque les récompenses attribuées à ces deux métriques sont identiques. En mettant l'accent sur la récompense liée au SCR_{marché}, le modèle parvient à mieux appréhender cette complexité, ce qui favorise des améliorations significatives. Dans une optique de pilotage du résultat, il est donc pertinent d'ajuster les besoins des poids de la récompense en fonction des besoins.

Après avoir étudié les impacts d'une modification de la structure de récompense, la prochaine partie s'intéresse à une stratégie d'optimisation qui optimise la PVFP/SCR_{souscription} au lieu du couple PVFP/SCR_{marché}.

IV.4.3 Sensibilité sur la PVFP/SCR_{souscription}

L'objectif de cette analyse de sensibilité, après avoir examiné le comportement de la stratégie combinant PVFP et SCR_{marché}, est de procéder à une étude similaire avec le couple PVFP et SCR_{souscription} en le comparant à la stratégie *Fixed-Mix*. Les trois figures ci-dessous présentent en détail par classes d'actifs les stratégies d'allocations :

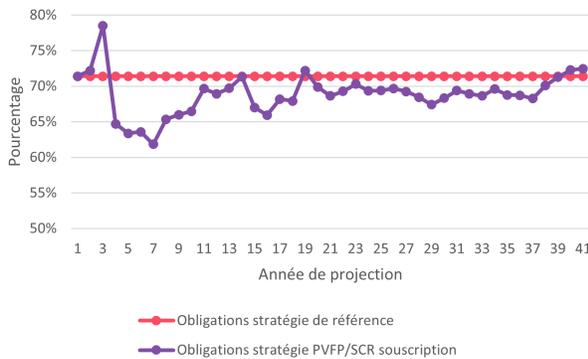


FIGURE IV.25 – Pourcentages des obligations au sein du portefeuille en fonction de la stratégie référence et PVFP/SCR_{souscription}

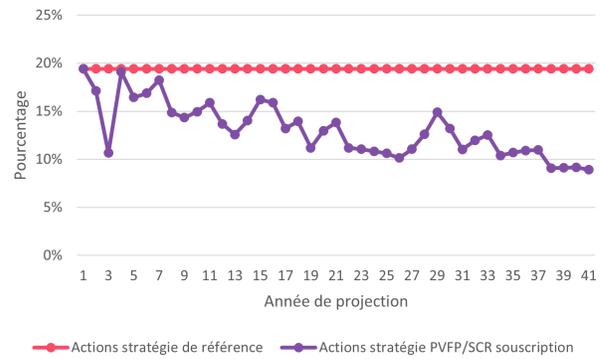


FIGURE IV.26 – Pourcentages des actions au sein du portefeuille en fonction de la stratégie référence et PVFP/SCR_{souscription}

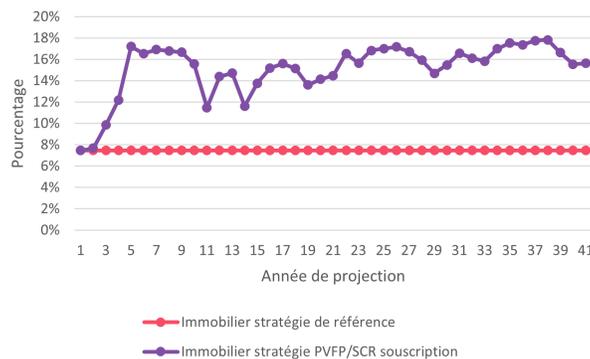


FIGURE IV.27 – Pourcentages de l'immobilier au sein du portefeuille en fonction de la stratégie référence et PVFP/SCR_{souscription}

Contrairement à la stratégie qui cible simultanément la PVFP/SCR_{marché}, la stratégie axée sur l'optimisation de la PVFP/SCR_{souscription} ne modifie que légèrement sa composition obligataire, maintenant une allocation similaire à celle observée dans le scénario de référence. En revanche, la part de l'immobilier dans le portefeuille explose passant de 8 % en moyenne à 15 %, tandis que la proportion d'actions

subit une baisse notable. Une première hypothèse concernant cette stratégie est qu'elle pourrait entraîner une détérioration de la PVFP par rapport à la stratégie qui optimisait la PVFP/SCR_{marché}. En effet, le modèle n'exploite plus pleinement l'avantage de l'intégration de nouvelles obligations offrant des taux d'intérêt plus élevés.

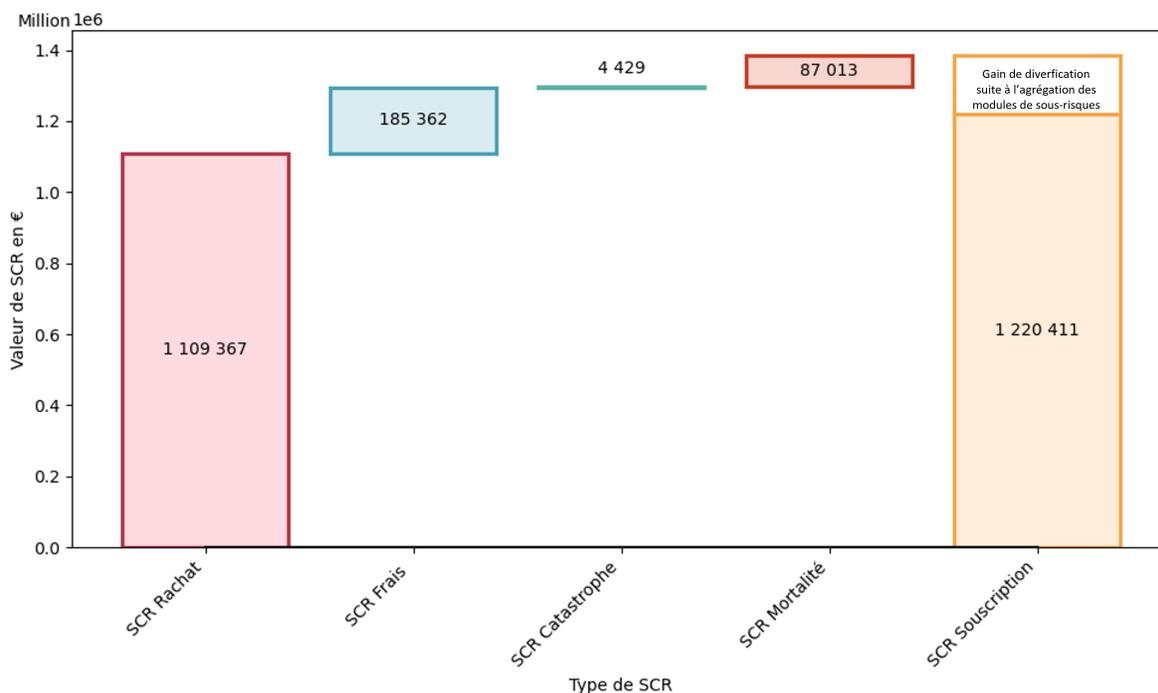
Métriques	Référence	PVFP/SCR _{souscription}
PVFP	3 449 110	3 488 505
SCR	2 672 117	2 718 537
SCR Marché	1 955 249	2 034 109
SCR Souscription	1 235 838	1 220 411

TABLE IV.6 – Comparaison des métriques entre les deux stratégies

Métriques	Écart relatif
PVFP	+1 %
SCR	+2 %
SCR Marché	+4 %
SCR Souscription	-1 %

TABLE IV.7 – Écart relatif des métriques par rapport à la stratégie de référence

Les deux tableaux ci-dessus (IV.6 et IV.7) présentent les comparaisons entre les métriques étudiées. Une première remarque est que l'objectif du modèle qui était de diminuer le SCR_{souscription} et d'augmenter la PVFP est rempli. En effet, la PVFP est beaucoup moins importante que dans les autres stratégies d'optimisation malgré la réalisation de PVL actions en début de projection à cause d'une part moins importante d'obligations en portefeuille profitant du contexte de taux élevés. Une diminution de -1 % du SCR_{souscription} est observée pour la stratégie optimisée par rapport à celle de référence. Pour mieux comprendre cette diminution, il est intéressant de se pencher sur la décomposition du SCR_{souscription} avec la figure suivante :



Composante du SCR	SCR Rachat	SCR Frais	SCR Cat	SCR Mortalité
Écart en € et en %	-17 031 (-1,5%)	1 672 (0,9%)	89 (2%)	4 381 (5,3%)

TABLE IV.8 – Décomposition du SCR souscription et écart en % de la stratégie PVFP/SCR_{souscription}

La figure ci-dessus (IV.8) illustre la décomposition du SCR_{souscription} pour la stratégie qui optimise la PVFP/SCR_{souscription}. Dans les deux cas, en scénario de référence et en stratégie optimisée, le SCR_{souscription} est grandement dominé par le SCR_{rachat} et plus particulièrement à cause du SCR_{rachat}^{massif} comme le montre la figure ci-dessous IV.28. Effectivement, en début de projection, la liquidation d'un volume important de l'encours du portefeuille expose à la réalisation de MVL obligataire et à une perte d'opportunité tout au long de la projection. Cela résulte du fait que l'assureur est limité en termes de volume dans son investissement, notamment dans des nouvelles obligations à des niveaux de taux d'intérêt plus élevés permettant ainsi d'améliorer son résultat financier.

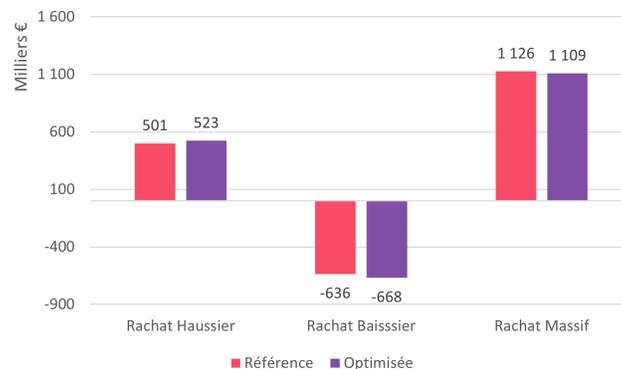


FIGURE IV.28 – Le SCR_{rachat} pour les deux stratégies

L'amélioration du SCR_{souscription} s'est donc faite par le biais d'une diminution du SCR_{rachat}^{massif}. Dans la stratégie optimisée, la vente de PVL sur les actions joue un rôle clé en contrebalançant les MVL importantes sur les obligations, résultant du choc de rachat massif. Ce surplus de produits financiers (IV.29) permet de mieux servir les assurés comme le montre la figure IV.30 et de les retenir plus longtemps dans la projection au sein du portefeuille. Par un effet de volume, cette stratégie augmente les profits financiers, réduisant ainsi le delta de PVFP sur le choc et, par conséquent, diminuant le SCR_{souscription}.

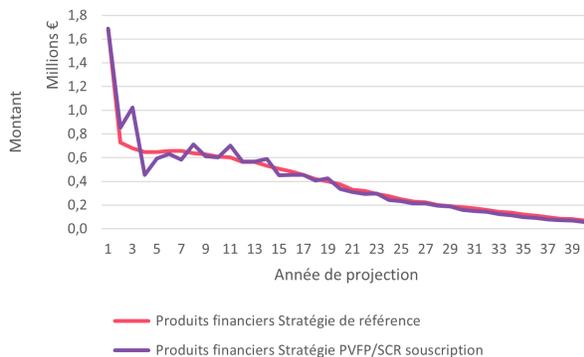


FIGURE IV.29 – Produits financiers dans le scénario de rachat massif pour les deux stratégies d’allocations

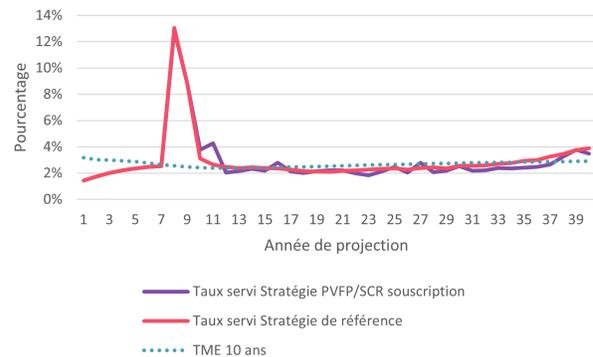


FIGURE IV.30 – TME, Taux servi des deux stratégies d’optimisation lors du scénario de rachat massif

La sensibilité axée sur l’optimisation de la $PVFP/SCR_{souscription}$ atteint certes ses objectifs principaux, mais elle présente plusieurs limites. Parmi celles-ci, une dégradation du SCR de l’ordre de 2 % est observée. De plus, cette stratégie conduit à une allocation importante en immobilier, ce qui rend le portefeuille plus vulnérable face aux risques de liquidité.

Après avoir étudié les sensibilités liées aux paramètres du modèle, notamment à travers l’ajustement des hyperparamètres, la structure de récompense ainsi que la modification des métriques prises en compte, l’analyse se tourne désormais vers les sensibilités économiques.

IV.5 Analyse de sensibilités économiques

Cette section est dédiée aux sensibilités économiques visant à examiner la réactivité du modèle face à des scénarios économiques variés. Deux situations spécifiques seront étudiées : la première se concentre sur l'impact de PVL obligataires en situation initiale, tandis que la seconde explore les répercussions d'une chute brutale des taux d'intérêt. L'objectif est de déterminer la capacité d'adaptation du modèle à ces environnements économiques distincts.

IV.5.1 Impact des plus-values latentes obligataires sur le modèle

L'objectif de cette sensibilité est d'analyser l'effet PVL obligataires sur le modèle, dans un contexte où le portefeuille a bénéficié de taux d'intérêt élevés, permettant ainsi le renouvellement de son stock d'obligations. Ce renouvellement a conduit à une situation où les obligations détenues se trouvent en position de PVL. Pour étudier cet impact, l'ensemble des autres inputs du modèle (courbe des taux, tables d'actifs, tables de passifs, etc...) sont conservés. La première étape implique l'évaluation du niveau de PVL sur les obligations du portefeuille. Pour ce faire, en se basant sur la situation économique au 31/12/2021 un benchmark auprès de différents assureurs a été réalisé, les PVL initiales sont alors définies comme suit :

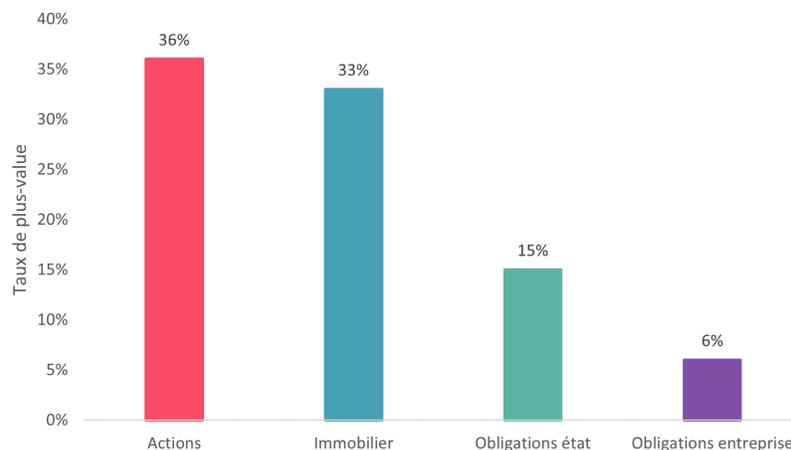


FIGURE IV.31 – Taux de PMVL initiales dans une situation initiale de PVL obligataires

La situation illustrée par le graphique ci-dessus IV.31 associée à un contexte de taux élevés constitue un environnement économique idéal par le fait que l'assureur peut chercher des produits financiers au travers des différentes classes d'actifs. Le portefeuille est effectivement en situation de PVL sur les obligations. D'un point de vue capital réglementaire, voici la décomposition du SCR donné par la stratégie de référence sans optimisation :

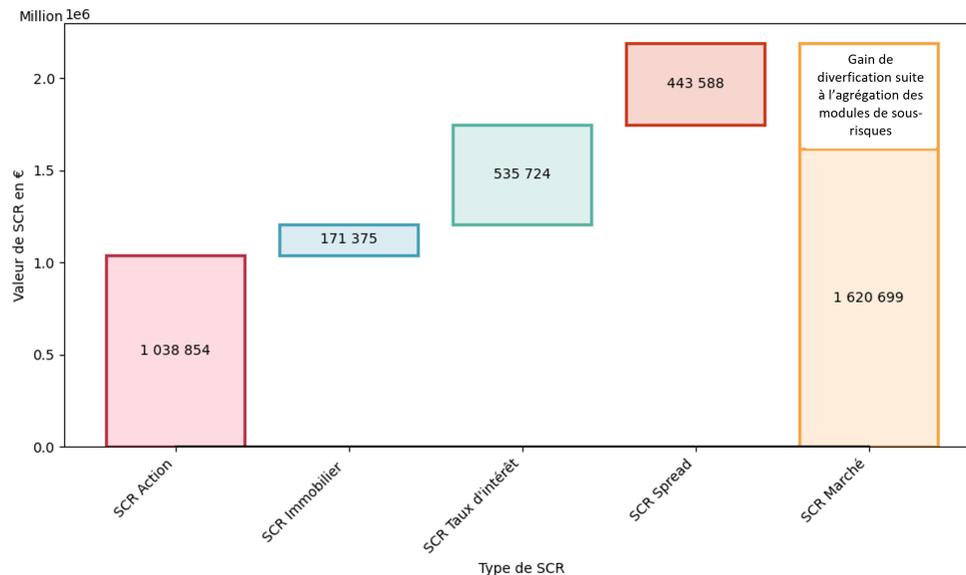


FIGURE IV.32 – Décomposition du SCR de marché pour la stratégie de référence

La décomposition du $SCR_{marché}$ (IV.32) issue de la stratégie de référence montre que le $SCR_{marché}$ est de nouveau dominé par SCR_{action} , il devra donc être l'objet d'une attention particulière lors de la stratégie d'optimisation. Dans la suite de cette sensibilité, la stratégie PVFP/ $SCR_{marché}$ est comparée de nouveau à celle de référence, mais aussi à la stratégie PVFP/ $SCR_{marché}$ en contexte de MVL obligataires de la partie IV.2.3. Cette démarche permet d'étudier comment le modèle a fait évoluer la stratégie d'allocation en prenant compte du changement d'environnement économique. Les trois figures ci-dessous présentent en détail par classes d'actifs les stratégies d'allocations :

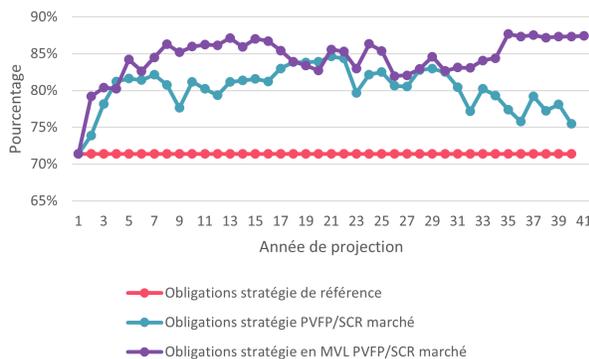


FIGURE IV.33 – Pourcentages des obligations au sein du portefeuille en fonction de la stratégie référence et PVFP/ $SCR_{marché}$

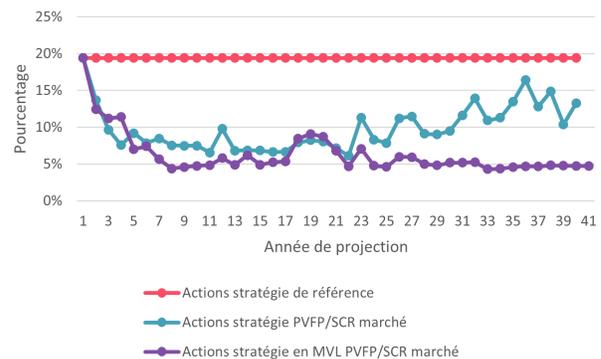


FIGURE IV.34 – Pourcentages des actions au sein du portefeuille en fonction de la stratégie référence et PVFP/ $SCR_{marché}$

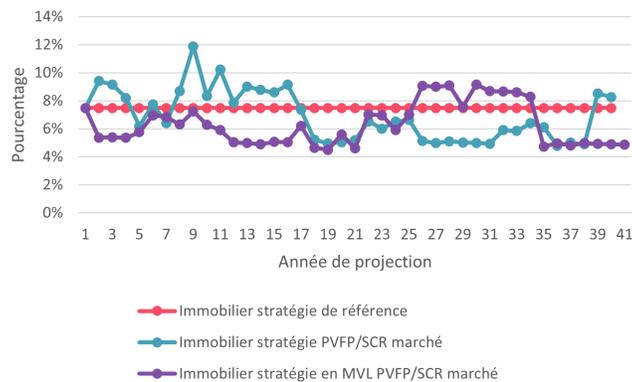


FIGURE IV.35 – Pourcentages de l’immobilier au sein du portefeuille en fonction de la stratégie de référence et PVFP/SCR_{marché}

Les résultats ci-dessus (IV.33, IV.34 et IV.35) représentent les différentes répartitions d’actifs au sein des allocations, mettant en lumière l’impact des PVL obligataires sur le portefeuille. Il est notable que la stratégie PVFP/SCR_{marché} ne privilégie pas autant les obligations comparativement à la stratégie MVL PVFP/SCR_{marché}. La stratégie PVFP/SCR_{marché} se désengage progressivement des obligations à partir de la 30ème année de projection, bénéficiant d’un double effet lié aux PVL obligataires et de taux élevés. De manière similaire, elle réalise massivement des PVL actions. Ce comportement suggère, d’une part, une diminution du SCR_{action}, le portefeuille étant proportionnellement moins exposé au choc des actions que la stratégie de référence, et d’autre part, une exposition accrue au risque de hausse des taux due à l’augmentation de la part obligataire. Dans ce contexte marqué par les PVL tant obligataires qu’actions, il est attendu que le modèle réussisse à générer davantage de produits financiers pour maximiser la PVFP.

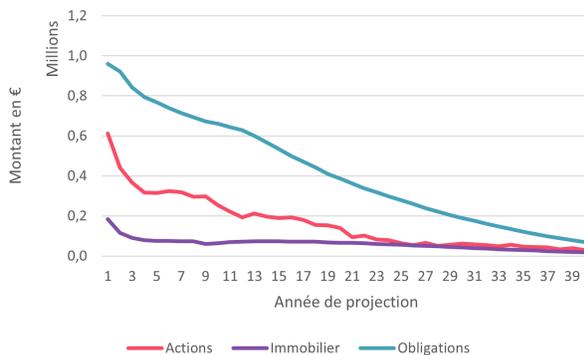


FIGURE IV.36 – Produits financiers générés pour les actifs de la stratégie référence

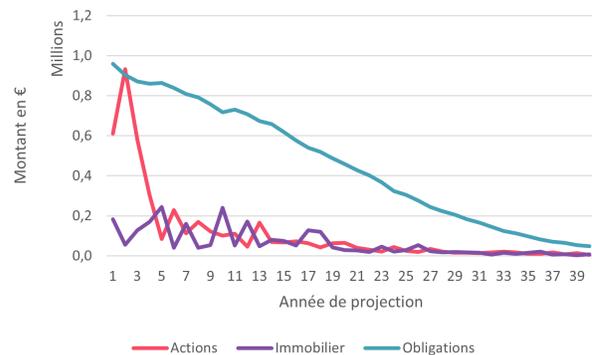
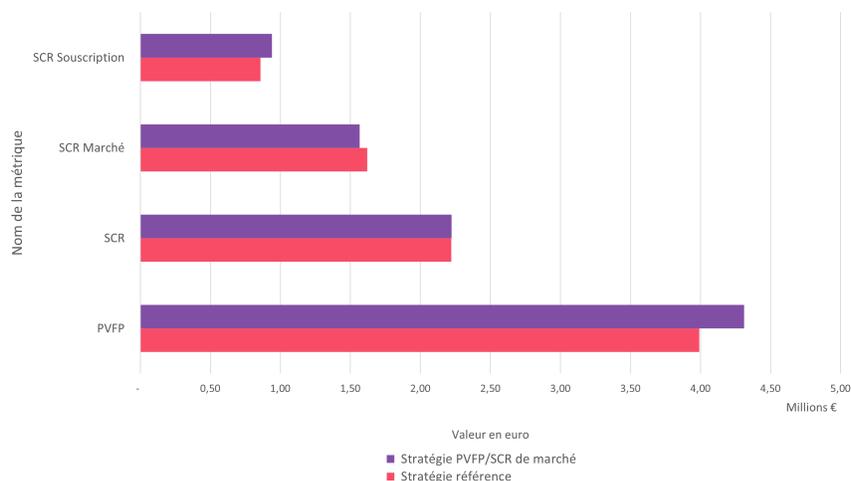


FIGURE IV.37 – Produits financiers générés pour les actifs de la stratégie PVFP/SCR_{marché}

Les deux figures ci-dessus (IV.36 et IV.37) appuient les remarques précédentes, notamment les PVL

massives actions réalisées par la stratégie PVFP/SCR_{marché} en début de projection ainsi qu'une poche obligataires plus importante permettant de générer plus de produits financiers. Il est intéressant de noter que les courbes de produits financiers entre l'action et l'immobilier s'entrecroisent démontrant la capacité du modèle à capter l'information sur le rendement financier des différents actifs.

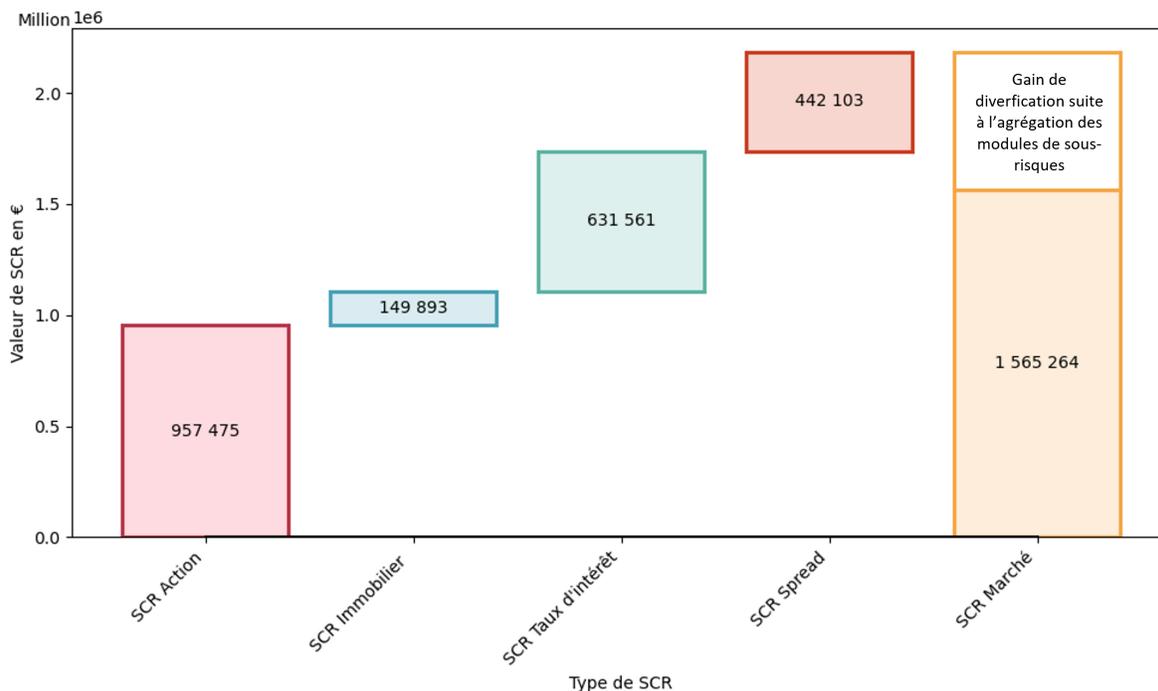


Métriques	Écart Relatif
PVFP	+8 %
SCR	0 %
SCR Marché	-4 %
SCR Soucription	+8 %

FIGURE IV.38 – Comparaison des métriques avec écart relatif par rapport au scénario de référence

La figure IV.38 présente les résultats obtenus par la stratégie. Les deux objectifs principaux de la stratégie, à savoir augmenter la PVFP et réduire le SCR_{marché}, sont effectivement réalisés. Le modèle tire profit de l'environnement économique favorable afin d'améliorer les métriques avec une hausse de +8 % pour la PVFP et une baisse de -4 % pour le SCR_{marché}. En effet, générant plus de produits financiers, cela permet au modèle de consolider sa position économique par le biais de la PVFP.

La figure IV.39 ci-dessous montre que l'optimisation du SCR_{marché} passe par une diminution du SCR_{action} et SCR_{immobilier}. Une diminution de la proportion en portefeuille des actions permet d'entraîner une amélioration du SCR_{action} grâce à un delta de PVFP moins important. Le même mécanisme s'applique à l'immobilier entraînant une diminution du SCR_{immobilier}.



Composante du SCR	SCR action	SCR immobilier	SCR taux	SCR spread
Écart en € et en %	-81 379 (-7,8%)	-21 482 (-12,5%)	95 837 (17,9%)	-1 485 (-0,3%)

FIGURE IV.39 – Décomposition du SCR de marché et écart en % de la stratégie PVFP/SCR_{marché} par rapport à la stratégie de référence

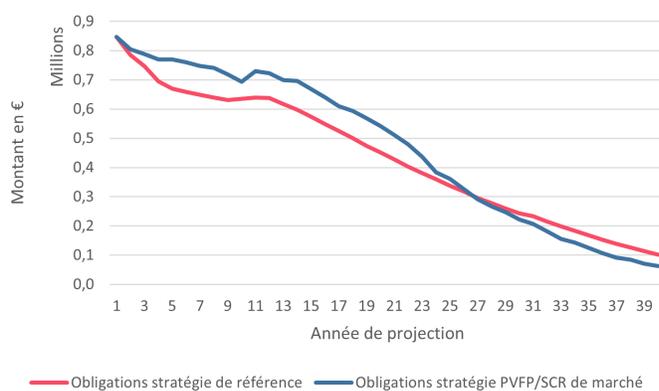


FIGURE IV.40 – Impact du choc hausse des taux sur les produits financiers obligataires des deux stratégies

Pour le SCR_{taux}^{up} , la sensibilité aux taux d'intérêt à la hausse persiste. Bien que la stratégie génère davantage de produits financiers obligataires comme le montre le graphique IV.40, le SCR_{taux}^{up} se dégrade. Cette dégradation s'explique par le fait que l'assiette de PVFP alimentée pour les produits financiers obligataires dans le modèle optimisé est plus conséquente que dans le modèle de référence.

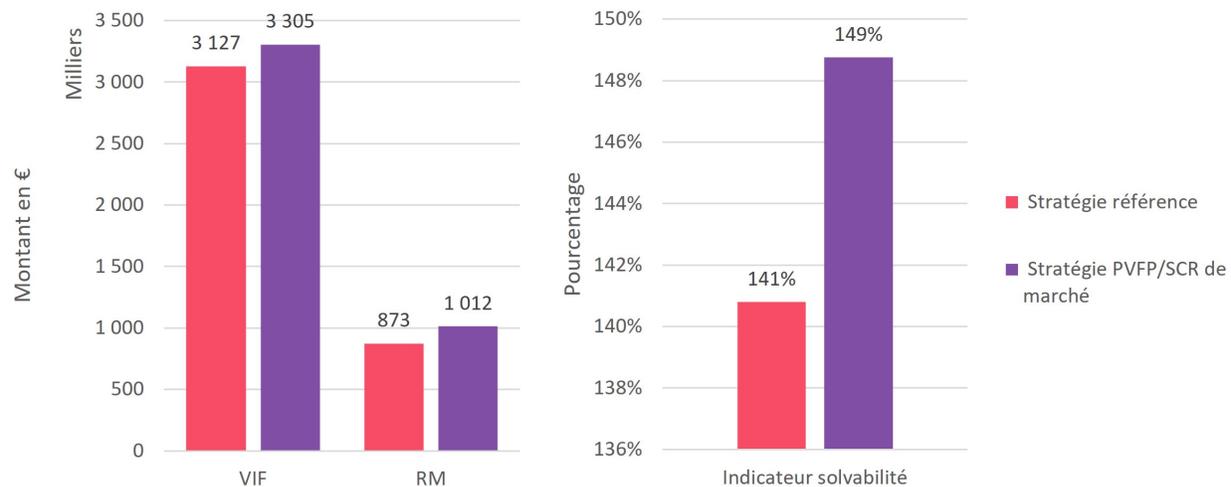


FIGURE IV.41 – Indicateur de solvabilité, RM ainsi que la VIF pour les deux stratégies

Le graphique présenté (IV.41) illustre comment la stratégie d'optimisation renforce la position réglementaire, faisant passer l'indicateur de solvabilité de 141 % à 149 %. Du fait que le SCR est stable, cette variation s'explique par l'augmentation significative de la PVFP malgré une dégradation du $SCR_{\text{subscription}}$ entraînant une hausse de la RM. La PVFP contribue à une amélioration de la VIF, soutenant donc la consolidation de l'indicateur de solvabilité. Le modèle permet donc, dans un contexte économique favorable, de dégager des optimisations financières significatives.

Après avoir examiné le comportement du modèle dans un contexte économique favorable, l'analyse se tourne désormais vers son adaptation dans un environnement plus hostile que celui du scénario central. L'étude continue par une analyse comportementale du modèle face à un choc de baisse de la courbe des taux.

IV.5.2 Impact d'un changement de la courbe des taux de -100bps

L'objectif de cette sensibilité est d'appliquer un choc sur la courbe des taux au 31/12/2022 utilisée dans le modèle pour examiner son impact sur la stratégie fournie par l'agent. Pour ce faire, la courbe des taux présentée dans le premier chapitre a été ajustée par un choc de -100bps. Ce choc a été choisi afin de pouvoir générer des impacts significatifs sur l'actif et le passif de l'assureur. Cela a permis de générer les résultats référencés au sein du graphique IV.42. Ensuite, en suivant la méthodologie proposée par l'EIOPA, les courbes correspondant aux scénarios de choc de baisse et de hausse des taux ont été reconstruites.

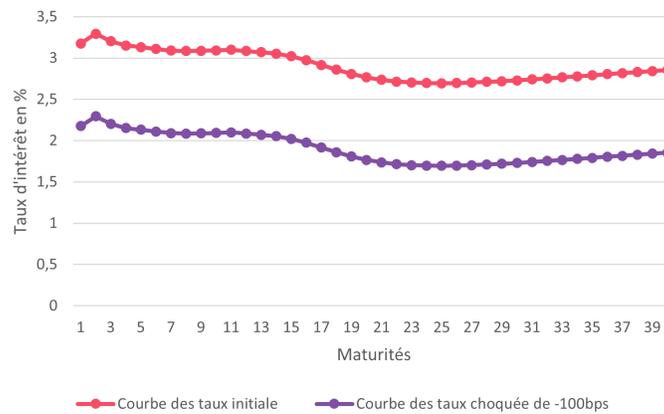


FIGURE IV.42 – Courbes des taux initiale et choquée de -100bps

Le but est donc de conserver l’environnement économique présenté dans le chapitre 1 en modifiant exclusivement la courbe des taux et par conséquent les inputs fournis par le GSE. Sous l’angle du capital réglementaire, voici comment se décompose le SCR de marché pour la stratégie de référence, avant toute optimisation :

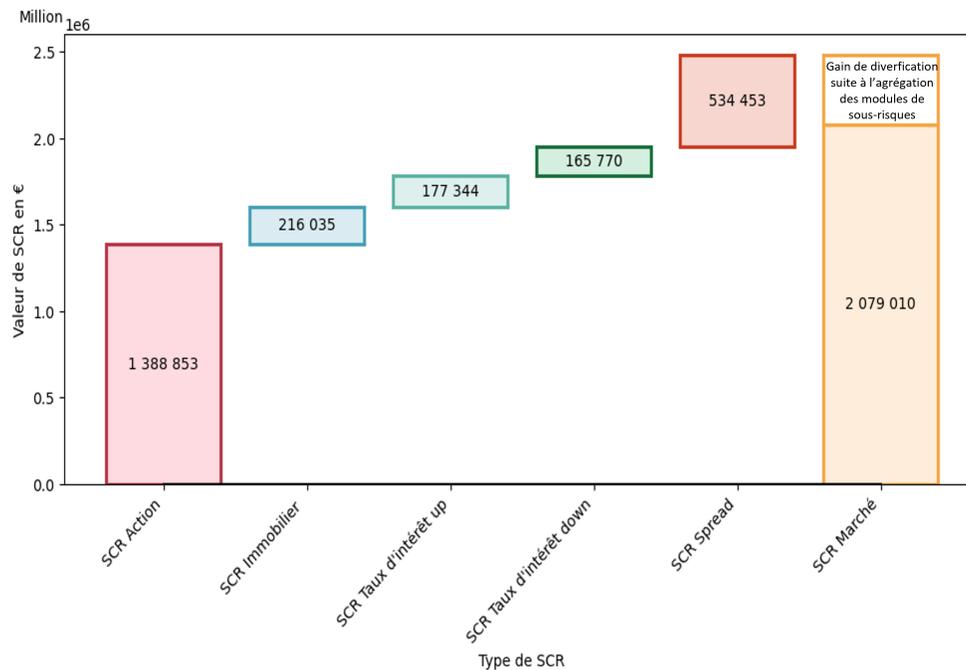


FIGURE IV.43 – Décomposition du SCR marché dans le scénario de référence

La figure ci-dessus IV.43 illustre le fait qu’encore une fois le $SCR_{marché}$ est dominé par le SCR_{action} .

Cependant, il est intéressant d'observer un premier impact du choc de la courbe des taux sur le SCR_{taux} . L'apparition d'une sensibilité aux baisses de taux et une nette réduction de la sensibilité aux hausses de taux est à noter. La sensibilité accrue aux baisses de taux découle de la stratégie de référence, qui tire son rendement des obligations pour servir le taux à ses assurés. Avec ce choc à la baisse, les nouvelles obligations acquises durant la projection en choc de taux à la baisse sont encore moins rémunératrices comparativement à la situation initiale avec la courbe des taux originale. Par conséquent, l'assureur éprouve plus de difficultés à pouvoir remplir ses engagements contractuels car pour rappel le TMG moyen net est de 0,77 % (cf.I.1).

Dans cette étude de sensibilité, la première étape consiste à présenter les pourcentages alloués à chaque catégorie d'actifs pour la stratégie *Fixed-Mix* et PVFP/ $SCR_{marché}$. Cette allocation sera comparée à celles des stratégies *Fixed-Mix* et PVFP/ $SCR_{marché}$ (OLD), telles que présenter avant l'application du choc sur la courbe des taux (IV.2.3).

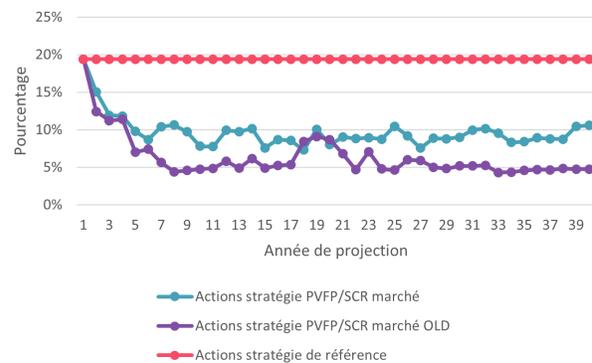
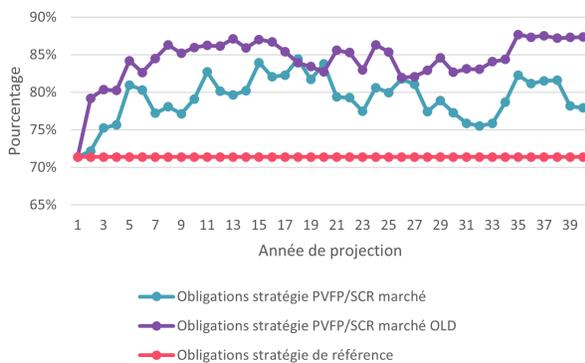


FIGURE IV.44 – Pourcentages des obligations au sein du portefeuille pour les trois stratégies

FIGURE IV.45 – Pourcentages des actions au sein du portefeuille pour les trois stratégies

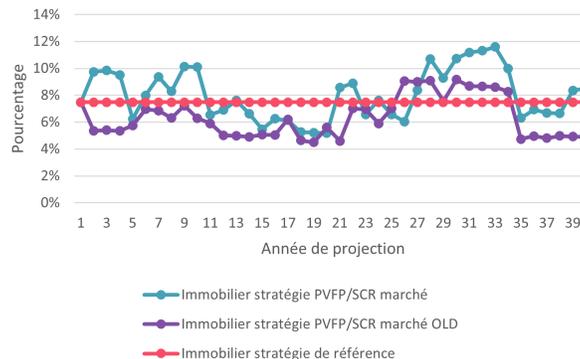


FIGURE IV.46 – Pourcentages de l'immobilier au sein du portefeuille pour les trois stratégies

Une première remarque est l'impact significatif du choc des taux sur l'allocation des obligations. En effet, la stratégie PVFP/SCR_{marché} ajuste sa part d'obligations en réduisant celle-ci par rapport à la version antérieure de la stratégie PVFP/SCR_{marché} OLD, en raison d'un rendement moins attractif des nouvelles obligations suite au choc des taux.

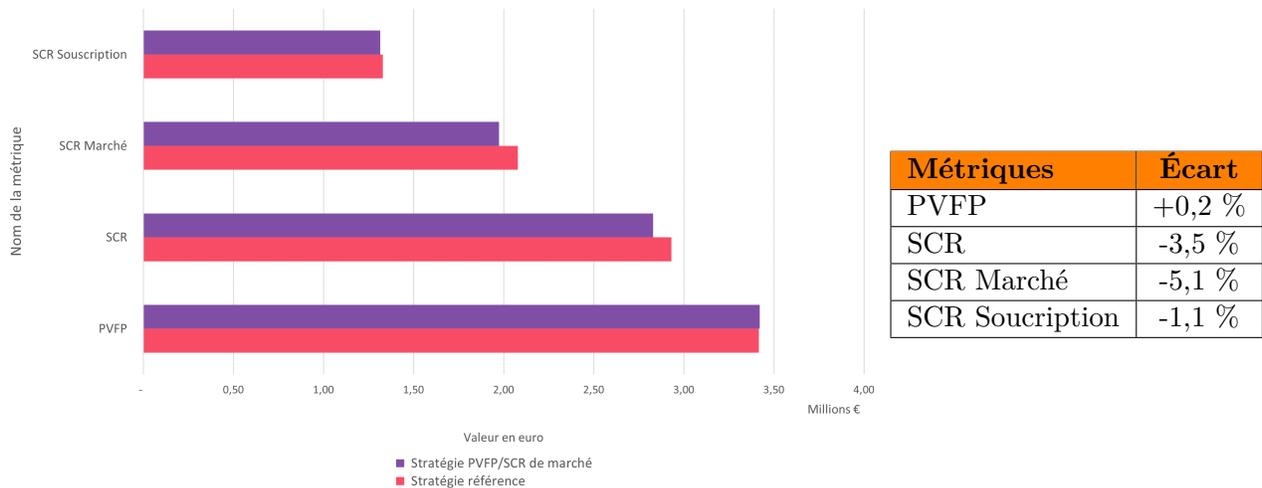


FIGURE IV.47 – Comparaison des métriques avec écart par rapport au scénario de référence

Dans cette sensibilité aux taux à la baisse, les objectifs du modèle d'optimiser le SCR_{marché} et la PVFP sont atteints. Toutefois, le modèle réagit de manière différente à ce nouvel environnement économique. En effet, cette fois-ci le modèle privilégie d'optimiser plus le SCR_{marché} (-5,1 %) en termes de volume que la PVFP (+0,2 %). Cela s'explique par le fait que l'optimisation de la PVFP peut être plus difficile à réaliser dans un environnement de taux plus bas, notamment car les obligations sont moins rémunératrices.

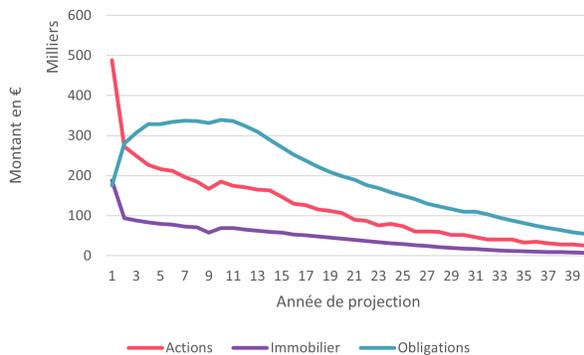


FIGURE IV.48 – Produits financiers générés pour les actifs de la stratégie référence

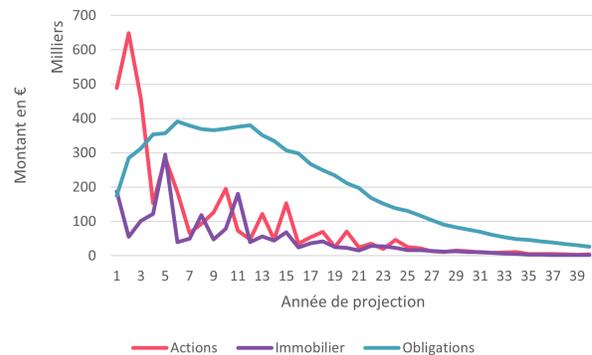
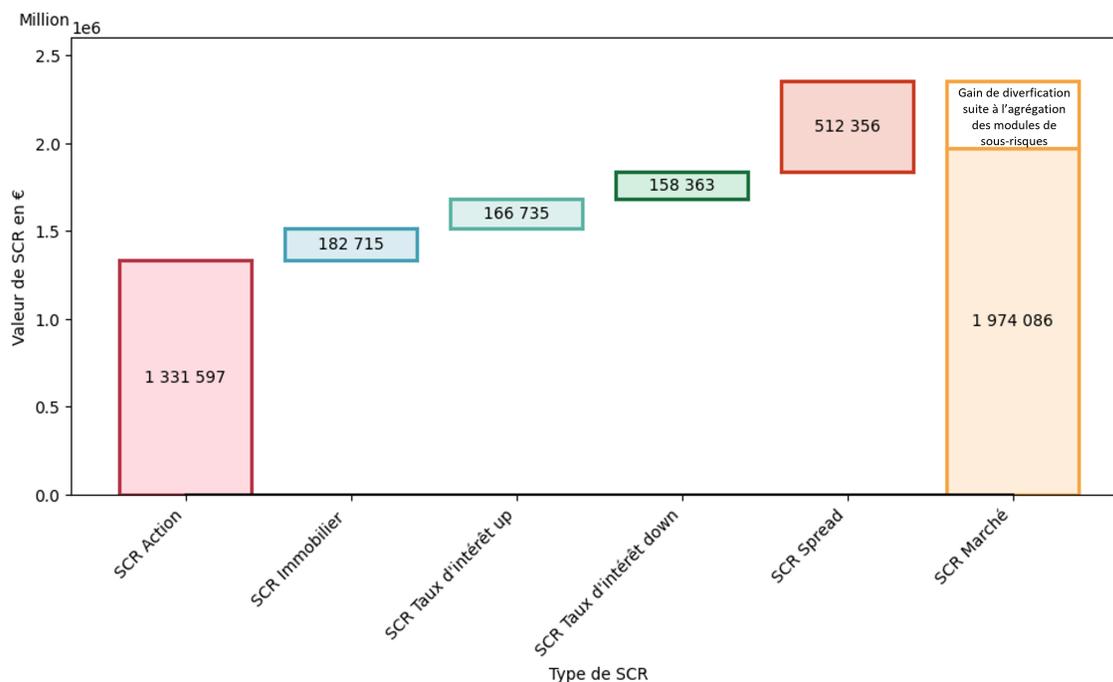


FIGURE IV.49 – Produits financiers générés pour les actifs de la stratégie PVFP/SCR_{marché}

Les deux graphiques (IV.49 et IV.48) présentant les produits financiers générés par les trois catégories d'actifs pour différentes stratégies illustrent l'impact du choc sur la courbe des taux. Un recul significatif des produits financiers provenant des obligations est observé en prenant comme élément de comparaison la sensibilité précédente, comme illustré par la référence IV.37. Ces résultats démontrent également la manière dont l'algorithme optimise la PVFP, avec une liquidation massive des PVL actions et l'acquisition de nouvelles obligations. Cependant, dans ce contexte, les obligations nouvellement acquises offrent un rendement moindre en raison de taux d'intérêt globalement plus bas d'où une amélioration faible de la PVFP.



Composante du SCR	SCR action	SCR immobilier	SCR taux up	SCR taux down	SCR spread
Écart en € et en %	-57 256 (-4,1 %)	-33 320 (-15,4 %)	-10 609 (-5,9 %)	-7 407 (-4,4 %)	-22 097 (-4,1%)

FIGURE IV.50 – Décomposition du SCR de marché et écart en % de la stratégie PVFP/SCR_{marché} par rapport à la stratégie de référence

La décomposition ci-dessus (IV.50) du SCR_{marché} montre que le modèle utilisé pour optimiser cette métrique procède à une consolidation de tous les sous-modules de ce SCR.

Premièrement, il désensibilise du portefeuille au SCR_{action} , notamment par une vente significative d'actions en début de projection, ce qui entraîne la réalisation de moins-values latentes sur actions. Néanmoins, en examinant l'écart de produit financier entre le scénario central et le scénario de choc sur les actions (IV.51), il apparaît que la stratégie PVFP/ $SCR_{marché}$ génère un écart moyen de produits financiers plus faible, à 99 k€ contre 107 k€ dans la stratégie de référence, donc un SCR_{action} moins élevé.

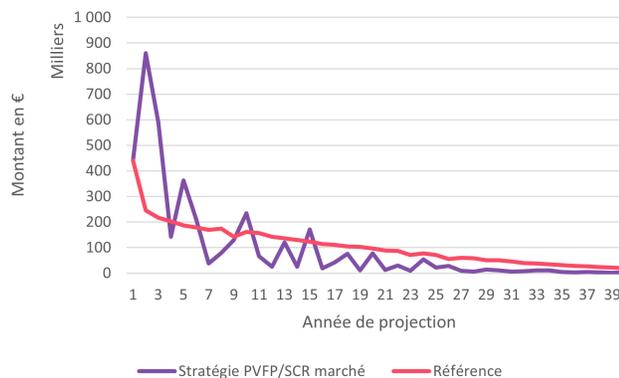


FIGURE IV.51 – Écart de produits financiers actions entre le scénario central et choc action pour les stratégies

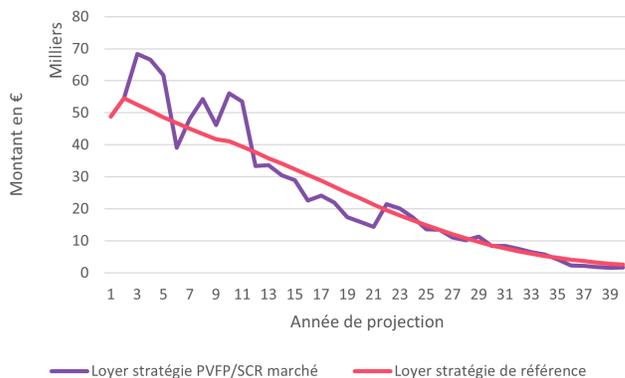


FIGURE IV.52 – Montant de loyer entre les deux stratégies
une meilleure protection contre le choc immobilier.

Le $SCR_{immobilier}$ diminue également, malgré une part en moyenne de 8 % contre 7,48 % pour la référence, entraînant ainsi un delta plus important de PVFP sur les plus-values latentes réalisées. Cette situation résulte d'un effet où le modèle privilégie les investissements immobiliers en début de projection, permettant ainsi de générer des montants de produits financiers plus conséquents, comme illustré par la figure IV.52. Cette stratégie contribue à accroître les produits financiers et offre

La réduction du SCR_{spread} est réalisée grâce au renouvellement des obligations. Dans la stratégie $SCR_{marché}$ décrite dans la section IV.2.1, possède une proportion équivalente d'obligations, cependant l'optimisation du SCR_{spread} dans cette stratégie est plus importante. Dans cette analyse de sensibilité, les nouvelles obligations affichent des taux d'intérêt plus bas, ce qui a pour effet de moins compenser efficacement le choc de spread, en comparaison avec les obligations de la courbe des taux sans choc.

Le renouvellement des obligations du portefeuille par des titres affichant des taux d'intérêt plus élevés induit un double effet bénéfique :

- Premièrement, dans le cas d'un choc de baisse des taux, cette stratégie contribue à générer des résultats financiers supérieurs. De plus, lors de ce choc, l'assureur n'est presque concerné que par le rachat structurel, car la probabilité de rachat conjoncturel, étant donné l'écart avec le TME, est au plus bas comme le montre la figure IV.54. Cela est dû au fait que le TME est presque au niveau du TMG moyen du portefeuille, à 0,77 % (cf. I.1). Ceci a pour effet d'améliorer le $SCR_{\text{taux}}^{\text{down}}$.
- Dans le cas d'un choc de hausse des taux, le même mécanisme permet de générer davantage de résultats financiers. De plus, comme le démontrent les deux figures suivantes (IV.53 et IV.55), le taux servi par la stratégie PVFP/ $SCR_{\text{marché}}$ dans le scénario de hausse des taux, sur la période de projection de la 9^{ème} à la 12^{ème} année, se maintient au-dessus du TME. Cette performance, comparée à la stratégie de référence, induit un effet de volume bénéfique, engendrant une augmentation des produits financiers par une rétention plus importante du portefeuille. Par ailleurs l'écart moyen du taux servi entre le scénario central et le choc à la hausse est plus important pour la stratégie de référence (+0,73 %) que pour la stratégie PVFP/ $SCR_{\text{marché}}$ (+0,66 %). Par conséquent, cela contribue à renforcer le $SCR_{\text{taux}}^{\text{up}}$.

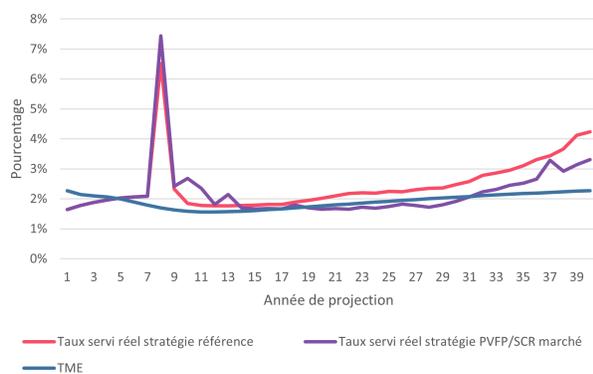


FIGURE IV.53 – Taux servi pour les deux stratégies en scénario central

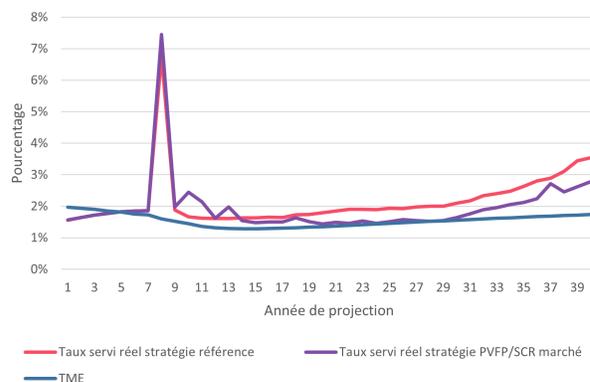


FIGURE IV.54 – Taux servi pour les deux stratégies en scénario choc de taux à la baisse

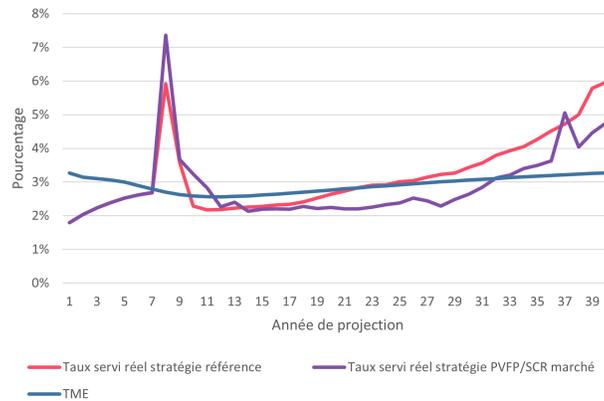


FIGURE IV.55 – Taux servi pour les deux stratégies en scénario choc de taux à la hausse

L'amélioration de l'indicateur de solvabilité, comme le montre la figure IV.56 ci-dessous de la stratégie PVFP/SCR_{marché} dans ce contexte de sensibilité de taux à la baisse, est attribuable à deux facteurs. En premier lieu, une diminution du SCR_{souscription} entraînant ainsi une baisse de la RM. De plus, le SCR global est également amélioré, ce qui permet, par effet combiné, de renforcer l'indicateur de solvabilité.

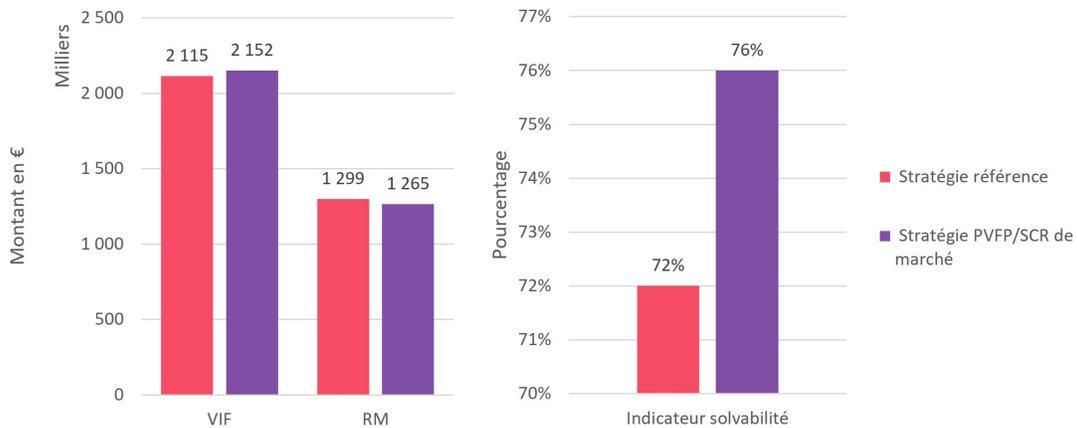


FIGURE IV.56 – Indicateur de solvabilité, RM ainsi que la VIF pour les deux stratégies

Le modèle a réussi donc à s'adapter à l'impact d'un choc de -100bps sur la courbe des taux. Cependant, un choc à la baisse dans cette analyse de sensibilité a inversé la proportion d'amélioration des métriques, favorisant davantage le SCR_{marché} que la PVFP par rapport à la stratégie PVFP/SCR_{marché} sans courbe choquée IV.2.3. Cette situation découle du contexte économique provoqué par une courbe des taux choquée, qui réduit les possibilités pour l'assureur d'améliorer sa PVFP. Ce choc se répercute sur le système de récompense du modèle, entraînant un écart peu important vis-à-vis de la situation initiale. La récompense est alors moindre par rapport à la récompense attribuée à l'amélioration du

$SCR_{marché}$. En effet, pour le $SCR_{marché}$, le modèle génère des écarts plus importants, ce qui se traduit par une récompense plus conséquente. C'est pour cette raison que la stratégie finale favorise nettement l'amélioration du $SCR_{marché}$ par rapport à la PVFP.

Conclusion

Cette étude a pour but de proposer à l'assureur une alternative pour sa stratégie d'allocation d'actifs en employant le *reinforcement learning*. L'approche, dans un contexte de taux élevés, a pour objectif d'optimiser son rendement financier tout en assurant la conformité avec les exigences réglementaires. Le *reinforcement learning* par l'algorithme DDPG fournit un outil flexible dans le pilotage de sa stratégie d'allocation d'actifs en adéquation avec ses besoins. Pour ce faire, l'étude se concentre sur quatre stratégies d'allocation d'actifs afin d'examiner la performance et le comportement de l'algorithme, appliquer sur les données au sein du modèle ALM :

- La stratégie PVFP (Maximiser la PVFP)
- La stratégie $SCR_{\text{marché}}$ (Minimiser le $SCR_{\text{marché}}$)
- La stratégie PVFP/ $SCR_{\text{marché}}$ (Maximiser la PVFP et minimiser le $SCR_{\text{marché}}$)
- La stratégie TRA/Richesse latente (Maximiser le TRA et maximiser la richesse latente)

Dans un contexte de taux élevés, l'algorithme démontre sa pertinence en augmentant significativement la PVFP, avec une hausse de 13 % pour la stratégie PVFP. La stratégie PVFP/ $SCR_{\text{marché}}$ enregistre une augmentation de 12 % de la PVFP et une réduction de 2 % du SCR de marché. De plus, pour évaluer la polyvalence du modèle, plusieurs sensibilités, tant au niveau du modèle que sur le plan économique, ont été conduites. Dans le scénario de sensibilité impliquant un choc à la baisse de la courbe des taux de -100 bps, l'algorithme a réussi à proposer une allocation permettant de diminuer le SCR de marché de 5 % pour la stratégie PVFP/ $SCR_{\text{marché}}$. De même, pour la sensibilité de PVL obligataires, l'algorithme parvient à tirer profit d'un environnement économique favorable, en augmentant la PVFP de 8 % et en réduisant le SCR de marché de 5 %

Les résultats obtenus démontrent que l'approche par *reinforcement learning* procure au modèle, toujours selon l'environnement économique et les métriques d'optimisations choisies, une réelle capacité d'adaptation. Cette approche permet de dégager des montants plus ou moins importants en fonction des différentes métriques ($SCR_{\text{marché}}$, PVFP, etc...) et de tirer profit des conditions du marché. Cependant, un point de vigilance a été soulevé notamment au travers des différentes sensibilités de

modèle concernant le processus de paramétrage. L'univers de projection, le poids, le nombre d'épisodes, ainsi que la structure de récompense exercent une influence significative sur les résultats. Cela exige une attention minutieuse lors de l'implémentation du modèle ainsi que les contrôles sur les propositions d'allocations. Une approche similaire à celle d'une recherche exhaustive de type "*grid search*" pour identifier les paramètres optimaux permet de tester la robustesse du paramétrage. En outre, étant donné que la durée d'entraînement peut s'avérer conséquente, dans une perspective d'industrialisation, il est envisageable de calibrer le modèle uniquement lors d'un trimestre de production. Par la suite, le modèle ainsi entraîné peut être utilisé pour les trimestres suivants, à condition qu'aucun changement majeur de l'environnement économique ne soit observé. Un changement majeur de l'environnement économique se réfère à une situation économique à laquelle le modèle n'a jamais été confronté auparavant.

Le modèle présente certaines limites qu'il convient de souligner. Une des principales restrictions réside dans le nombre de métriques prises en compte pour offrir une analyse plus complète. Par ailleurs, le modèle révèle un comportement parfois erratique, comme en témoigne la variation abrupte du pourcentage d'actions d'une année à l'autre dans certaines allocations, indiquant un lissage moins important du résultat financier que dans la stratégie *Fixed-Mix*. Cela est lié aux hypothèses de variations d'actifs à chaque pas de temps qui ont été prises, ce qui, dans un contexte réel, pourrait engendrer des coûts de transactions significatifs. En outre, le modèle ne prend pas en compte la liquidité du marché, notamment pour l'immobilier.

L'objectif principal de ce mémoire est de proposer une alternative à la stratégie *Fixed-Mix* actuellement en place. Au vu des résultats obtenus, la stratégie optimisant le couple la PVFP/SCR_{marché} se présente comme un candidat idéal. Cette stratégie a démontré sa robustesse et sa capacité d'adaptation à divers environnements économiques ainsi qu'à différents paramétrages, soulignant son potentiel en tant que solution d'optimisation efficace.

Bien que ces résultats soient encourageants, il serait intéressant de se projeter sur de potentielles améliorations de ce modèle au-delà du cadre actuel. Notamment, il offre un potentiel significatif pour être intégré dans des processus stratégiques de type ORSA permettant ainsi à l'assureur de piloter son business plan de manière agile. De plus, l'approche actuelle pourrait être enrichie par l'inclusion de métriques liées au passif, tel que le taux servi aux assurés permettant ainsi d'avoir une vision plus holistique, autant côté actif que passif, de la stratégie d'allocation d'actifs.

Bibliographie

- ALAEDDINE, F. (2012). « Allocation stratégique d'actifs et ALM pour les régimes de retraite ». Mémoire d'actuariat.
- CALIXTE, H. (2010). « Etude de la convergence du générateur de scénarios économiques servant aux calculs de l'European Embedded Value pour les Variable Annuities ». Mémoire de Master.
- DISERBEAU, Y. (2019). « Une proposition d'accélérateurs pour la mise en place d'une couverture indicielle des risques météo-sensibles ». Mémoire d'actuariat.
- KIM, Daehwan et Anthony M. SANTOMERO (1988). « Risk in Banking and Capital Regulation ». In : *Journal of Finance*.
- LÉGIFRANCE (2023). « Code des assurances. Article L.331-3 ». In : Consulté le 06/11/2023.
- LEIBOWITZ, Martin L. (1992). « Asset Performance and Surplus Control : A Dual-Shortfall Approach ». In : *The Journal of Portfolio Management*.
- MACAULAY, Frederick R. (1938). *The Movements of Interest Rates. Bond Yields and Stock Prices in the United States since 1856*. New York : National Bureau of Economic Research.
- MARKOWITZ, Harry (1952). « Portfolio Selection ». In : *The Journal of Finance*.
- MASSON, M. (2021). « Leviers sur la solvabilité en assurance vie dans un contexte de taux bas ». Mémoire d'actuariat.
- MNIH V, et al. (2015). « Human-level control through deep reinforcement learning ». Research paper.
- NAHON, J. (2022). « Apport de la garantie fidélité en assurance vie ». Mémoire d'actuariat.
- PEROLD, Andre F. et William F. SHARPE (1988). « Dynamic Strategies for Asset Allocation ». In : *Financial Analysts Journal*.
- REDINGTON, F.M. (1952). « Review of the Principles of Life-Office Valuations ». In : *Journal of the Institute of Actuaries (1886-1994)*.
- RODRIGUES FONTOURA, A (2020). *A deep reinforcement learning approach to asset-liability management*. <https://eic.cefet-rj.br/ppcic/wp-content/uploads/2020/07/21-Alan-Rodrigues-Fontoura.pdf>.
- SHARPE, William F. et Lawrence G. TINT (1990). « Liabilities : A New Approach ». In : *Journal of Portfolio Management*.
- SILVER, David (2014). *Deterministic Policy Gradient Algorithms*.
- SUTTON, R et A BARTO (2017). *Reinforcement Learning, An Introduction, Second edition*. MIT Press.
-

TAMBRUN, H. (2020). « Allocation stratégique d'actifs en épargne dans le cadre d'une remontée rapide des taux d'intérêt ». Mémoire d'actuariat.

TIMOTHY P, et al. (2016). « Continuous control with deep reinforcement learning ». Research paper.

WENG, L (2018). « A (Long) Peek into Reinforcement Learning ». In : URL : <https://lilianweng.github.io/posts/2018-02-19-rl-overview/>.

Annexes

A.1 Matrices de corrélation $SCR_{marché}$

$SCR_{marché}$	Intérêt	Équité	Immobilier	Spread	Change	Concentration
Intérêt	1.00	0.50	0.50	0.50	0.25	0.00
	1.00	0.00	0.00	0.00	0.25	0.00
Équité	0.50	1.00	0.75	0.75	0.25	0.00
	0.00	1.00	0.75	0.75	0.25	0.00
Immobilier	0.50	0.75	1.00	0.50	0.25	0.00
	0.00	0.75	1.00	0.50	0.25	0.00
Spread	0.50	0.75	0.50	1.00	0.25	0.00
	0.00	0.75	0.50	1.00	0.25	0.00
Change	0.25	0.25	0.25	0.25	1.00	0.00
	0.25	0.25	0.25	0.25	1.00	0.00
Concentration	0.00	0.00	0.00	0.00	0.00	1.00
	0.00	0.00	0.00	0.00	0.00	1.00

TABLE .9 – Matrices de corrélation fusionnées pour les sous-modules du SCR de marché dans les scénarios *Down* et *Up*. Les valeurs en bleu clair indiquent des valeurs qui correspondent à un choc des taux à la baisse. Les valeurs en bleu foncé indiquent des valeurs qui correspondent à un choc des taux à la hausse.

A.2 Définitions martingales et mouvement brownien

Martingales

Soit $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ un espace de probabilité filtré.

Un processus aléatoire $(X_t)_{t \geq 0}$ est une martingale si :

1. $E[|X_t|] < \infty$,
2. X_t est \mathcal{F}_t -adapté,
3. $\forall s < t, \quad E[X_t | \mathcal{F}_s] = X_s$.

Mouvement Brownien

Le **mouvement brownien**, également connu sous le nom de processus de Wiener, est un exemple de processus stochastique qui joue un rôle central en mathématiques financières et en théorie des probabilités. Il est défini par les propriétés suivantes :

1. $W_0 = 0$,
2. Les accroissements $W_t - W_s$ sont indépendants pour $s < t$,
3. Les accroissements $W_t - W_s$ suivent une distribution normale avec une moyenne de 0 et une variance de $t - s$,
4. W_t est continu presque partout.

A.3 Portefeuille pour la frontière de Markowitz

Paramètre	Valeur
Nombre de portefeuilles simulés	10,000
Noms des actifs	Asset 1 ; Asset 2
Rendements moyens	0.1 ; 0.2
Matrice de covariance	[0.05, 0.02] ; [0.02, 0.08]

TABLE .10 – Paramètres du portefeuille pour la frontière efficiente

A.4 L'algorithme *Q-learning*

Algorithm 1 Algorithme Q-learning

```

1: Initialiser  $Q(s, a)$  arbitrairement pour tout  $s \in S, a \in A(s)$ 
2: Initialiser  $s$ 
3: for chaque épisode do
4:    $s \leftarrow$  état initial de l'épisode
5:   while  $s$  n'est pas un état terminal do
6:     Choisir  $a$  à partir de  $s$  en utilisant une politique dérivée de  $Q$  (par exemple,  $\epsilon$ -greedy)
7:     Prendre l'action  $a$ , observer la récompense  $r$  et l'état suivant  $s'$ 
8:      $Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$ 
9:      $s \leftarrow s'$ 
10:  end while
11: end for

```

A.5 Démonstration du Théorème du Gradient de la Politique

1. Définition de $J(\theta)$ en utilisant Q :

$$J(\theta) = \mathbb{E}_{\pi_\theta}[G_t] = \mathbb{E}_{\pi_\theta}[Q_\pi(s, a)] \quad (1)$$

2. Application de la décomposition d'une espérance en somme pour $J(\theta)$:

$$J(\theta) = \sum_{s \in \mathcal{S}} \rho_\pi(s) \sum_{a \in \mathcal{A}} \pi_\theta(a|s) Q_\pi(s, a) \quad (2)$$

où $\rho_\pi(s)$ est la distribution stationnaire de la chaîne de Markov pour un état s sous la politique π .

3. Calcul du gradient de $J(\theta)$:

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} \rho_\pi(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q_\pi(s, a) \quad (3)$$

4. Réécriture en utilisant l'identité $\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s) (\nabla_\theta \ln \pi_\theta(a|s))$:

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} \rho_\pi(s) \sum_{a \in \mathcal{A}} \pi_\theta(a|s) (\nabla_\theta \ln \pi_\theta(a|s)) Q_\pi(s, a) \quad (4)$$

5. Réécriture sous forme d'espérance :

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \ln \pi_\theta(a|s) Q_\pi(s, a)] \quad (5)$$