



Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du  
Diplôme d'Actuaire EURIA  
et de l'admission à l'Institut des Actuares

2021

Par : CHÉRET Marine

Titre : Optimisation de l'étape de régression de l'approche LSMC pour le calcul d'un proxy du Best Estimate dans un modèle interne

Confidentialité : Non

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

**Membre présent du jury de l'Institut  
des Actuares :**

LE DILY Simon

HAMON Fabrice

Signature :

**Entreprise :**

Generali France

Signature :

**Membres présents du jury de l'EURIA : Directeur de mémoire en entreprise :**

DERRIEN Jean-Marc

BENNIS Abdelkrim

Signature :

**Autorisation de publication et de mise en ligne sur un site de diffusion  
de documents actuariels**

*(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise :

Signature du candidat :



## Résumé

En assurance vie, les interactions entre l'actif et le passif, créées par les options et garanties des contrats, obligent l'assureur à valoriser son passif de manière stochastique. Dans le cadre de l'utilisation d'un modèle interne, le calcul du besoin en capital (SCR) par projection de la distribution des fonds propres à un an nécessite alors la réalisation de simulations stochastiques imbriquées (Simulations dans les Simulations, SdS) : des simulations primaires projetant l'environnement économique à 1 an en monde réel puis des simulations secondaires pour valoriser le passif dans chacun des scénarios. Cela est parfois irréalisable de manière opérationnel dû à un problème de temps de calcul. Afin d'obtenir une forme paramétrique du *Provisions Best Estimate* à 1 an en fonction des facteurs de risque une des approches alternatives aux SdS est la méthode de *Least Square Monte Carlo* (LSMC). Cette méthode consiste à réaliser un jeu complet de simulations primaires et seulement quelques simulations secondaires pour la valorisation du BE, puis à effectuer une régression des valeurs du BE ainsi obtenues afin d'obtenir une fonction paramétrique du BE en fonction des facteurs de risques identifiés. L'étape de régression est usuellement faite par une régression polynomiale des facteurs de risque.

Cette étude a pour objectif de challenger cette régression par des modèles de Machine Learning. Pour cela les modèles de forêt aléatoire, *Extrême Gradient Boosting*, *Support Vector Regression* et *Multivariate Adaptive Regression Spline* ont été calibrés sur des valeurs de BE approchées par deux simulations de Monte Carlo antithétiques. Les résultats de prédiction des BE stressés et les performances des modèles sont ensuite comparés à ceux obtenus par la régression polynomiale actuellement en place. Enfin, le modèle le plus performant, sera utilisé pour calculer le SCR de la ligne d'activité (LoB) étudiée afin de comparer les résultats obtenus avec ceux actuels.

**Mots clefs:** ALM, Meilleure Estimation, LSMC, *Least Square Monte Carlo*, Modèle interne, SCR, Apprentissage Automatique, Régression.



## Abstract

In life insurance, the interactions between assets and liabilities, created by contract options and guarantees, require the insurer to value its liabilities stochastically. When using an internal model, the calculation of the Solvency Capital Requirement (SCR) by projecting the distribution of equity capital to one year, requires the performance of nested stochastic simulations : outer simulations projecting the economic environment to one year in the real world conditions and then inner simulations, to value the liabilities in each of the outer scenarios. This is sometimes not operationally feasible due to computational time issues. In order to obtain a parametric form of the one-year Best Estimate as a function of the risk factors, the alternative approach to those time consuming simulations is the Least Square Monte Carlo (LSMC) method. This method consists of carrying out a complete set of outer simulations and only a few inner simulations for the valuation of the BE, and then regressing the BE values. The goal here is to obtain a proxy function of the BE as a function of the identified risk factors. The regression step is usually done by a polynomial regression of the risk factors.

The aim of this study is to challenge this regression with Machine Learning models. For this purpose, the Random Forest, Extreme Gradient Boosting, Support Vector Regression and Multivariate Adaptive Regression Spline models were calibrated on BE values approximated by two antithetic Monte Carlo simulations. The prediction results of the stressed BE and the performance of the models are then compared to those obtained by the current polynomial regression. Finally, the best performing model will be used to calculate the SCR for the line of business studied. We will then compare the results obtained with those currently in place.

**Keywords:** ALM, Best Estimate, LSMC, Least Square Monte Carlo, Internal Model, Machine learning, Regression.



# Note de synthèse

## Calcul du capital de solvabilité requis en modèle interne

Solvabilité 2 impose aux assureurs européens un suivi constant de leurs risques et de leur solvabilité. Entre autre, la directive demande de calculer un montant de capital minimum à détenir pour que la probabilité de ruine économique à un horizon de 1 an de la compagnie d'assurance soit inférieure ou égale à 0.5% : le SCR. De plus, ce montant de SCR doit être suivi régulièrement tout au long de l'année. Dans le cadre d'un calcul par modèle interne, ce qui est le cas chez Generali, la valeur du SCR s'obtient par projection de la distribution des Fonds Propres, ou du passif *Best Estimate*, à 1 an via la formule suivante :

$$SCR_0 = FP_0 - q_{0.5\%}(D(1)FP_1)$$

avec

- $D(1) = \frac{1}{1+r}$ , le facteur d'actualisation, où r correspond au taux sans risque.
- $FP_1$  les Fonds Propres stressés à t=1 an.
- $q_{0.5\%}$  le quantile à 0.5%.

## Réalisation de simulations stochastiques imbriquées pour l'obtention de la distribution du passif à 1 an

Dû à la présence d'interactions entre l'actif et le passif créée par les options et garanties des contrats d'assurance vie, les provisions *Best Estimate*<sup>1</sup> se valorisent de manière stochastique, ce qui rend complexe l'obtention de la distribution des passifs et donc des Fonds Propres stressés à 1 an. En effet, l'approche la plus naturelle permettant d'obtenir la distribution complète des Fonds Propres stressés nécessite alors un calcul stochastique imbriqué afin de capter la valeur temps des options et garanties présentes dans les contrats. Cela implique l'exécution de plusieurs milliers de simulations primaires (monde réel) projetant les conditions économiques stressées à 1 an, suivi d'un grand nombre de simulations secondaires (risque neutre) afin de calculer, de manière stochastique, une valeur fiable de la provision *Best Estimate* liée à chacun des scénarios primaires. La figure suivante schématise cette approche.

---

1. Elles correspondent à l'engagement de l'assureur envers ses assurés

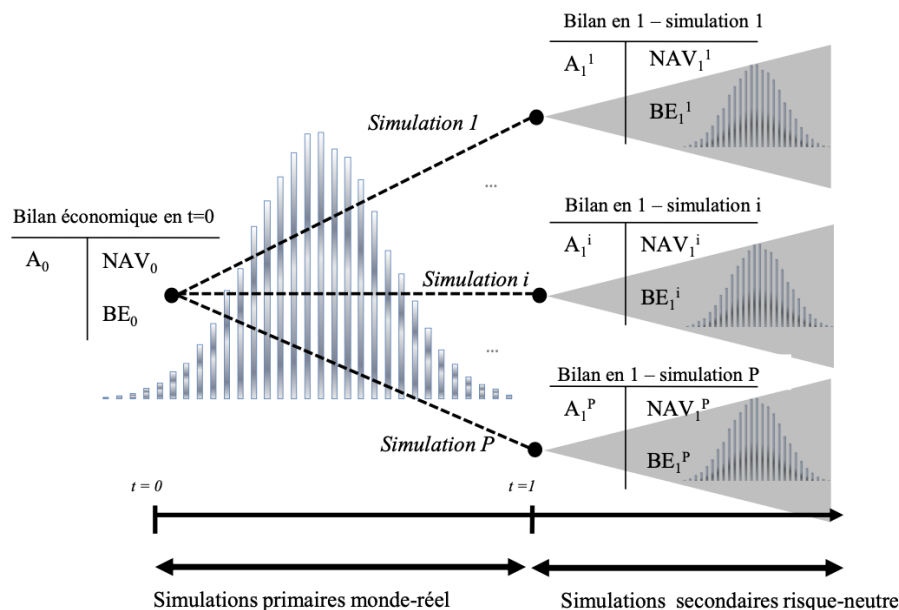


FIGURE 1 – Schématisation de l’approche des Simulations dans les Simulations  
 Source : Cours de modélisation et agrégation des risques

Cette approche stochastique imbriquée, nommée Simulation dans les Simulations, nécessite un temps de calcul déraisonnable voire infaisable opérationnellement tant la charge de calcul est importante.

### L’approche Least Square Monte Carlo

Différentes approches ont alors vu le jour afin de résoudre le problème de temps de calcul inhérent aux simulations dans les simulations, comme l’approche de *Replicating Portfolio*, l’approche *Curve Fitting* ou encore l’approche *Least Square Monte Carlo*. Parmi ces différentes approches, Generali a choisi de valoriser ses passifs d’assurance vie par une approche de Least Square Monte Carlo.

Le schéma ci-dessous permet de visualiser le fonctionnement de l’approche LSMC. Cette méthode consiste à calibrer une fonction paramétrique des passifs stressés à 1 an<sup>2</sup> à partir d’un nombre restreint de simulations. En effet, seulement 2 simulations secondaires antithétiques seront effectuées pour valoriser le *Best Estimate* dans chaque scénario primaire, puis une régression de ces valeurs peu fiable de *Best Estimate* en fonction des facteurs de risque sera réalisée. Cette fonction de régression permettra donc

2. NB : De manière opérationnelle dans le modèle interne de Generali, le passif n’est pas projeté à 1 an mais les chocs sont appliqués à  $t = 0^+$ , ce qui permet d’éviter de projeter l’environnement économique de 0 à 1 an.



d'obtenir une forme paramétrique du BE stressé en fonction des différents facteurs de risque.

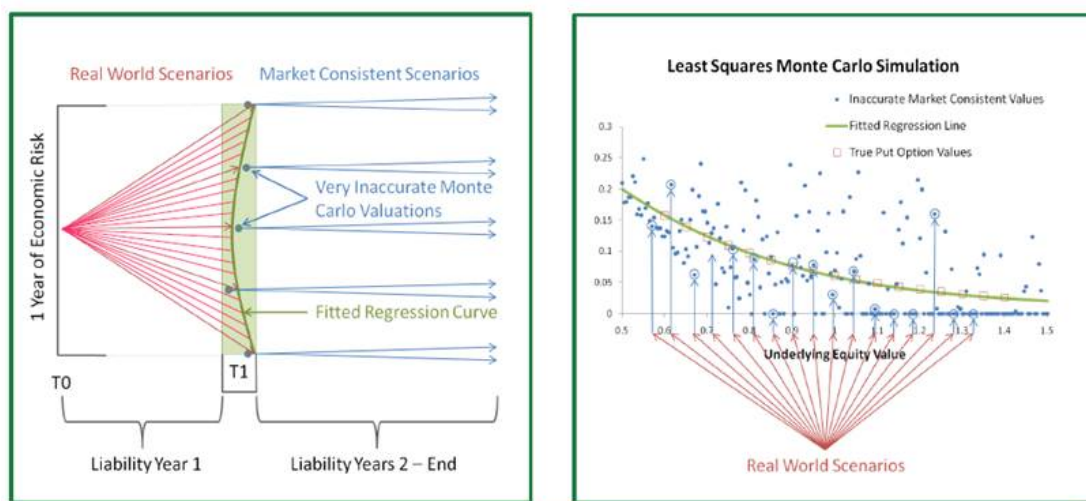


FIGURE 2 – Schématisation de l'approche Least Square Monte Carlo

Source : Documentation interne à Generali sur la méthodologie du modèle interne

### Régression effectuée au sein de l'approche LSMC

La régression effectuée au sein de l'approche LSMC, en vue d'obtenir une fonction paramétrique du BE en fonction des facteurs de risque, a été réalisée sur une base de données construite de la manière suivante : Un individu correspond à un scénario primaire, pour lequel on connaît la valeur des différents facteurs de risque ainsi que le *Best Estimate* correspondant obtenu par seulement deux simulations de Monte Carlo. Chez Generali, pour la ligne d'activité regroupant les contrats d'assurance vie contenant des interactions ALM (LoB unique), cette base contient 25 000 lignes, soit 25 000 scénarios primaires, et 14 facteurs de risque.

Actuellement dans le modèle interne, la régression réalisée dans cette approche est une régression polynomiale, contenant des termes croisés de plusieurs facteurs de risque. Les termes de cette régression polynomiale sont rigoureusement sélectionnés par un algorithme de sélection de termes *Adaptative Forward Stepwise*, basé sur le critère AIC<sup>3</sup>, afin d'en maximiser les performances.

L'objectif de ce mémoire est de chercher à mettre à l'épreuve cette régression polynomiale améliorée par des algorithmes de *Machine Learning*.

3. L'AIC, *Akaike Information Criterion*, est une mesure de la qualité d'un modèle statistique, basée sur la vraisemblance et pénalisant la complexité du modèle.

## Régression par *Machine Learning* en vue d'obtenir une forme paramétrique du BE par LSMC

À partir de cette même base de données, quatre modèles de *Machine Learning* ont été calibrés en vue de challenger la performance et/ou le temps d'élaboration de la forme paramétrique actuelle du BE stressé :

- *Random Forest* (RF)
- *Extrême Gradient Boosting* (XGB)
- *Support Vector Machine* adapté à la régression (SVR)
- *Multivariate Adaptive Regression Splines* (MARS)

Les performances des 4 modèles ainsi construits ont ensuite été calculées sur une base de validation, contenant quant à elle 125 valeurs de *Best Estimate* fiables en fonction des 14 facteurs de risque. Afin d'obtenir des valeurs précises de *Best Estimate* pour chacun des 125 scénarios, 1000 simulations secondaires Monte Carlo ont été réalisées, amenant à un budget de 125 000 scénarios pour la construction de cette base de validation.

Voici les performances de ces différents modèles sur la prédiction du *Best Estimate* stressé de la ligne d'activité d'assurance vie<sup>4</sup>, après avoir optimiser le choix de leurs hyper-paramètres par validation croisée :

Modèles	RMSE	WRE	$\epsilon \max$ <sup>5</sup>	$R^2$
<b>Polynôme actuel</b>	<b>150.5</b>	<b>0.19%</b>	<b>0.5%</b>	<b>99.8%</b>
<b>Random Forest</b>	311	0.36%	1.4%	99.1%
<b>XGBoost</b>	268	0.30%	1.3%	99.4%
<b>SVM</b>	208	0.24%	0.85%	99.6%
<b>MARS</b>	<b>146.7</b>	<b>0.173%</b>	<b>0.66%</b>	<b>99.8%</b>

FIGURE 3 – Métriques calculées sur la base de validation pour chaque modèle.

Il ressort de ce tableau récapitulatif que le modèle MARS surpasse les autres modèles de *Machine Learning* testés. De plus, c'est le seul des quatre modèles qui arrive à surpasser légèrement les performances du polynôme actuel. Ce modèle est également très intéressant en terme de temps de calcul, avec une durée de *fitting*<sup>6</sup> inférieure à 1 minute.

Nous avons alors cherché à connaître l'impact qu'aurait l'utilisation de la fonction de régression de MARS pour l'obtention de la distribution complète du *Best Estimate* stressé dans le calcul du SCR.

4. LoB unique d'assurance vie, regroupant tous les contrats présentant des interactions ALM.

6. Temps nécessaire à la calibration du modèle, une fois les hyper-paramètres déterminés.

## Impact d'un changement de forme paramétrique pour l'estimation du *Best Estimate* sur le SCR

Le capital de solvabilité requis en vision modèle interne se calcule par projection de la distribution complète des Fonds Propres à partir de l'approximation suivante :

$$SCR_0 = FP_0 - q_{0.5\%}(FP_{stress\ 0+})$$

En effet, dans cette étude les chocs sont appliqués en  $t = 0^+$ , ce qui évite de projeter l'environnement économique à 1 an.

Les Fonds Propres peuvent se déduire des autres postes du bilan ainsi

$$FP_{0+} \approx Actif_{stress\ 0+} - BE_{stress\ 0+} - OBSI_{stress\ 0+}$$

En remplaçant  $BE_{stress\ 0+}$  par sa forme paramétrique polynomiale, puis par la fonction de régression obtenue par MARS, nous avons pu calculer l'impact sur le SCR du changement de forme paramétrique pour l'approximation de BE.

$SCR_{MARS}$	$SCR_{POLY}$	Écart	%
1 941 M	2 071 M	- 130 M	<b>-6.3%</b>

FIGURE 4 – Comparaison des SCR en fonction des fonctions répliques du BE utilisées

Malgré le fait que les performances des deux fonctions sur la base de validation soient proches, nous observons un écart significatif de - 130 millions d'euros sur le capital de solvabilité requis en changeant de forme paramétrique pour l'obtention de la distribution du *Best Estimate*.

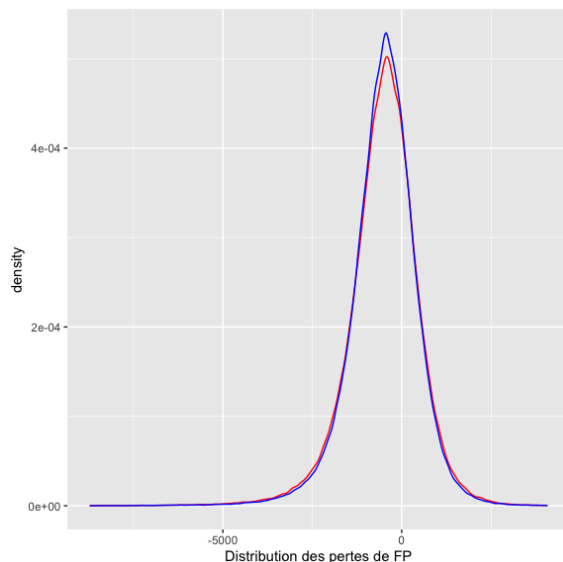


FIGURE 5 – Distribution des pertes de Fonds Propres stressés. En rouge est représentée la distribution par le polynôme et en bleu par MARS.

## Analyse de l'écart observé

En s'intéressant à la distribution complète des pertes de Fonds Propres avec chacune des approches, il est remarquable que les queues de distribution via l'approche polynomiale sont plus épaisses. De plus, les pertes de Fonds Propres obtenues par l'approche polynomiale présentent quelques valeurs beaucoup plus extrêmes qu'avec MARS.

Ceci peut être dû aux scénarios *Out of Range*, c'est à dire ceux qui sortent du champs de calibration des modèles. En effet, les deux modèles ont été construits à partir d'une base d'apprentissage sur laquelle les différents facteurs de risque étaient compris entre le quantile 0.1% et le quantile 99.9% de leur distribution. Mais, lorsque nous appliquons ces fonctions à l'ensemble de la *Risk Factor Scénario Set* (RFSS)<sup>7</sup>, certains des scénarios sortent du champs d'apprentissage. Or, nous n'avons aucune maîtrise des fonctions paramétriques sur ces zones là.

Le modèle MARS a l'avantage de rester linéaire en dehors des intervalles sur lesquels il a appris, ce qui permet de rester en adéquation avec la tendance observée sur les données d'apprentissage, et d'éviter d'atteindre des valeurs de *Best Estimate* stressé trop explosives. Quant à lui, le polynôme est totalement incontrôlable et peut atteindre des extrêmes importants à cause des différentes puissances des facteurs de risque. En plus d'obtenir des valeurs explosives, le polynôme peut changer totalement d'allure en dehors des intervalles d'apprentissage, comme c'est le cas sur la variable *IdxBond*, représentant le risque de dégradation du marché obligataire. Le risque obligataire est le facteur le plus impactant sur la valeur du *Best Estimate* avec une corrélation très positive, or comme le montre la figure suivante, le polynôme prédit une corrélation extrêmement négative en dehors des zones d'apprentissage.

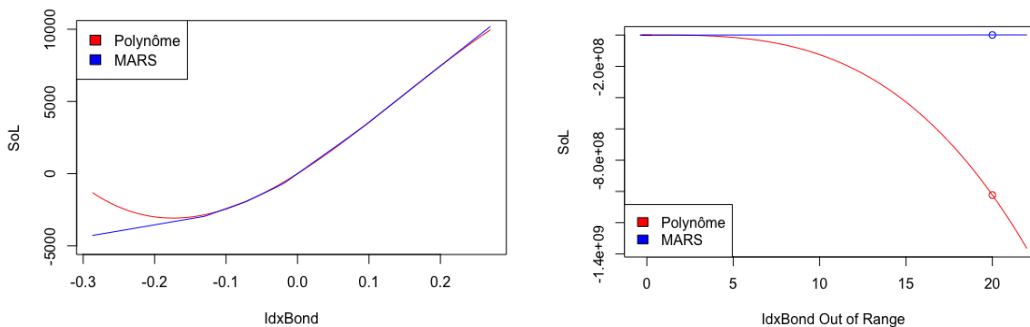


FIGURE 6 – Réaction des proxys en dehors des régions d'apprentissage de *IdxBond*

3.2% des scénarios de la RFSS ont au moins un facteur de risque qui sort de la zone d'apprentissage des fonctions paramétriques, et 1.05% des scénarios de la RFSS

7. Ensemble de  $2^8$  scénarios permettant d'obtenir la distribution complète des passifs.

ont l'un des 5 facteurs de risque les plus impactants (à savoir IdxBond, Idx\_Eq, dVA, IR\_PC1\_EUR et Idx\_Pr) en dehors de la zone d'apprentissage.

Afin de réduire l'erreur liée à ces scénarios *Out of Range*, il pourrait être pertinent d'élargir l'intervalle sur lequel les modèles sont construits en passant par exemple du quantile 0.05% au quantile 99.95% de la distribution de chacun des facteurs de risque. Il est néanmoins possible que cela détériore les performances des modèles sur les zones centrales.

# Executive summary

## Internal Model Solvency Capital Requirement Calculation

Solvency 2 requires European insurers to constantly monitor their risks and solvency. Among other things, the directive requires the calculation of a minimum amount of capital to be held so that the probability of economic ruin over a one-year horizon is less than or equal to 0.5 % : the SCR. Moreover, this amount of SCR must be monitored throughout the year. In the context of a calculation by internal model, which is the case at Generali, the value of the SCR is obtained by projecting the distribution of the Equity, or the liabilities Best Estimate, at 1 year via the following formula :

$$SCR_0 = FP_0 - q_{0.5\%}(D(1)FP_1)$$

with  $D(1)$  the discount factor,  $FP_1$  the stressed Equity at  $t=1$  year.

## Running nested stochastic simulations to obtain the distribution of liabilities at 1 year

However, due to the presence of interactions between the assets and liabilities created by the options and guarantees of life insurance contracts, the provisions Best Estimate<sup>8</sup> are valued stochastically, which makes it complex to obtain the distribution of liabilities and therefore stressed equity at 1 year. Indeed, the most natural approach to obtain the complete distribution of the Stressed Equity requires a nested stochastic calculation to capture the time value of the options and guarantees present in the contracts. This involves running several thousand primary (real-world) simulations projecting stressed economic conditions to 1 year, followed by a large number of secondary (risk-neutral) simulations in order to stochastically calculate a reliable value of the provision linked to each of the primary scenarios. The following figure illustrates this approach.

---

8. They correspond to the insurer's commitment to its policyholders

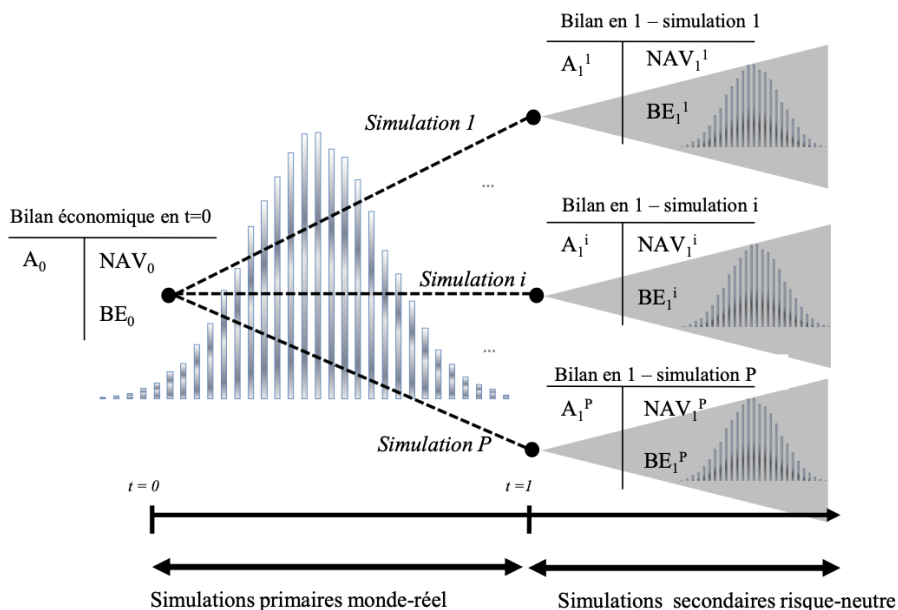


FIGURE 7 – Schema of the nested stochastic approach *Source : Risk modelling and aggregation course*

This nested stochastic approach, requires an unreasonable or even operationally unfeasible calculation time as the calculation load is too high.

### The Least Square Monte Carlo approach

Various approaches were then developed to solve the problem of calculation time inherent in simulations, such as the Replicating Portfolio approach, the Curve Fitting approach or the Least Square Monte Carlo approach. Among these different approaches, Generali has chosen to value its life insurance liabilities using a Least Square Monte Carlo approach.

The diagram below shows how the LSMC approach works. This method consists of calibrating a parametric function of the stressed liabilities at 1 year NB : Operationally, in Generali’s internal model, the liabilities are not projected to 1 year but the shocks are applied at  $t = 0^+$ , which avoids projecting the economic environment from 0 to 1 year from a limited number of simulations. Indeed, only 2 antithetical secondary simulations will be carried out to value the Best Estimate in each primary scenario, then a regression of these unreliable values of Best Estimate according to the risk factors will be performed. This regression function will therefore provide a parametric form of the stressed BE as a function of the different risk factors.

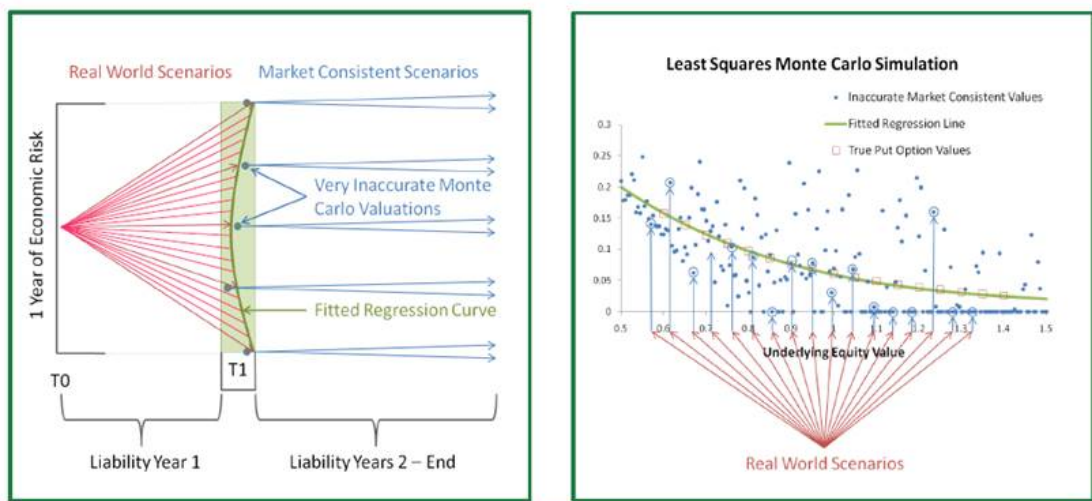


FIGURE 8 – Schematic representation of the Least Square Monte Carlo approach  
 Source : Generali internal documentation on the methodology of the internal model

### Regression within the LSMC approach

The regression carried out within the LSMC approach, in order to obtain a parametric function of the BE as a function of the risk factors, is carried out on a database constructed in the following way : An individual corresponds to a primary scenario, for which the value of the different risk factors is known, as well as the corresponding Best Estimate obtained by only two Monte Carlo simulations. At Generali, this database contains 25,000 lines, i.e. 25,000 primary scenarios, and 17 risk factors.

Currently in the internal model, the regression performed in this approach is a polynomial regression, containing cross terms of several risk factors. The terms of this polynomial regression are rigorously selected by a term selection algorithm, based on the AIC criteria, in order to maximise its performance.

The objective of this thesis is to challenge this polynomial regression improved by Machine Learning algorithms.

### Machine Learning regression to obtain a parametric form of the BE by LSMC

From this same database, four Machine Learning models were calibrated in order to challenge the performance and/or the development time of the current parametric form of the stressed BE :

- Random Forest (RF)
- Extreme Gradient Boosting (XGB)



- Support Vector Machine adapted to regression (SVR)
- Multivariate Adaptive Regression Spline (MARS)

The performances of the 4 models thus constructed were then calculated on a validation basis, containing 125 reliable values of Best Estimate according to the 14 risk factors. In order to obtain precise values of Best Estimate for each of the 125 scenarios, 1000 secondary Monte Carlo simulations were carried out, leading to a budget of 125,000 scenarios for the construction of this validation base. Here are the performances of these different models, after optimising the choice of their hyper-parameters by cross-validation :

Models	RMSE	WRE	$\epsilon \max^9$	$R^2$
<b>Polynomial</b>	<b>150.5</b>	<b>0.19%</b>	<b>0.5%</b>	<b>99.8%</b>
<b>Random Forest</b>	311	0.36%	1.4%	99.1%
<b>XGBoost</b>	268	0.30%	1.3%	99.4%
<b>SVR</b>	208	0.24%	0.85%	99.6%
<b>MARS</b>	146.7	0.173%	0.66%	99.8%

FIGURE 9 – Metrics calculated on the basis of validation for each model

It is clear from this summary table that the MARS model outperforms the other Machine Learning models tested. Moreover, it is the only one of the four models that manages to slightly outperform the current polynomial. This model is also very interesting in terms of calculation time, with a fitting time of less than 1 minute.

We then wanted to know the impact on the SCR that would create the use of the regression function of the MARS model to obtain the complete distribution of the stressed BE.

### Impact of a change of parametric form for the estimation of the Best Estimate on the SCR

The Solvency Capital Requirement in the internal model view is calculated by projecting the complete distribution of the Equity by the following approximation :

$$SCR_0 = FP_0 - q_{0.5\%}(FP_{stress\ 0+})$$

In fact, in this study the shocks are applied in  $t = 0^+$ , which avoids projecting the economic environment to 1 year.

Equity can be deduced from the other balance sheet items as follows

$$FP_{0+} \approx Assets_{stress\ 0+} - BE_{stress\ 0+} - OBSI_{stress\ 0+}$$

By replacing  $BE_{stress\ 0+}$  by its polynomial parametric form, then by the regression function obtained by MARS, we were able to calculate the impact on the SCR of the change of parametric form for the BE approximation.

$SCR_{MARS}$	$SCR_{POLY}$	Gap	%
1,941 M	2,071 M	- 130 M	<b>-6.3%</b>

FIGURE 10 – Comparison of SCRs according to the replicative functions of the EB used

Despite the fact that the performances of the two functions on the validation basis are close, we observe a significant difference of - €130 million on the Solvency Capital Requirement by changing the parametric form for obtaining the Best Estimate distribution

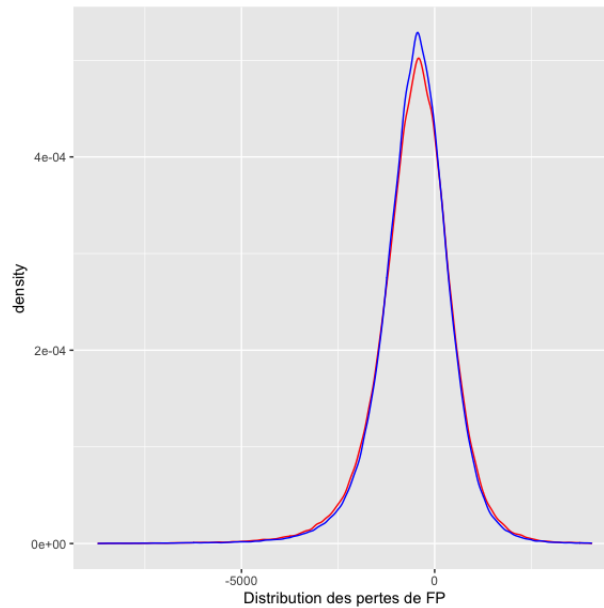


FIGURE 11 – Distribution of the Losses of Stranded Equity  
In red is represented the polynomial distribution and in blue by MARS

### Analysis of the observed gap

Looking at the full distribution of equity losses with each of the approaches, it is remarkable that the distribution tails via the polynomial approach are thicker. In addition, the equity losses obtained with the polynomial approach show some much more explosive values than with MARS.

The reason for this may be the Out of Range scenarios, i.e. those that fall outside the models' calibration fields. Indeed, both models were built from a training base on which the different risk factors were between the 0.1% and 99.9% quantile of their distribution. But, when we apply these functions to the whole Risk Factor Scenario Set (RFSS), some of the scenarios fall outside the learning field. We have no control over the parametric functions in these areas.

The MARS model has the advantage of remaining linear outside the intervals on which it has learned, which makes it possible to remain in line with the trend observed on the training data, and avoids reaching values of Best Estimate that are too extreme. The polynomial, on the other hand, is totally uncontrollable and can explode due to the different powers of the risk factors. In addition to obtaining explosive values, the polynomial can change its shape completely outside of the learning intervals, as it is the case for the IdxBond variable, which is also the one that has the greatest impact on the SCR value.

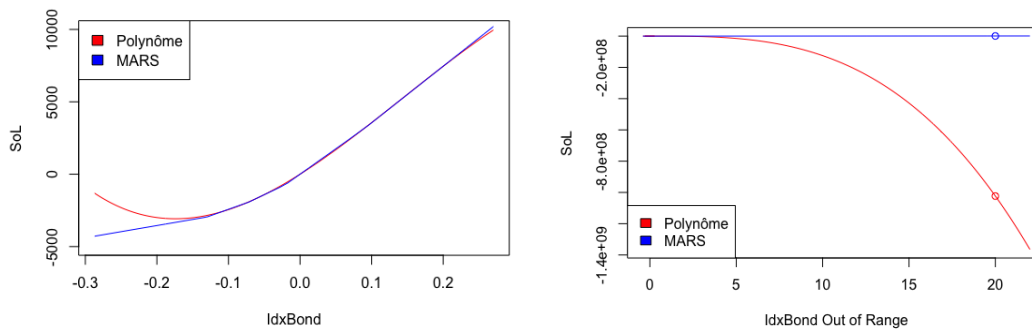


FIGURE 12 – Reaction of proxies outside the IdxBond learning regions

3.2% of the RFSS scenarios have at least one risk factor that falls outside the learning region of the parametric functions, and 1.05% of the RFSS scenarios have the 5 most impactful risk factors (i.e. IdxBond, Idx\_Eq, dVA, IR\_PC1\_EUR and Idx\_Pr) outside the learning region.

In order to reduce the error linked to these Out of Range scenarios, it could be relevant to widen the interval on which the models are built by passing for example from the 0.05% quantile to the 99.95% quantile. But it is possible that this would deteriorate the performance of the models on the central zones.

# Remerciements

Je souhaite remercier toutes les personnes ayant contribué à la réalisation, ou à l'encadrement de ce mémoire.

Je remercie tout particulièrement mon tuteur d'entreprise, M. Abdelkrim Bennis, pour son implication dans mes travaux de mémoire, son aide ainsi que sa disponibilité tout au long de mon mémoire, qui ont permis un bon encadrement et une bonne organisation générale.

Je tiens à remercier également Mme. Salimata Diop et M. Marco Longo, managers au sein de la direction des Risques de Generali, pour le suivi et l'intérêt porté à mes travaux.

Je remercie Mme Corinne Lefumat, tutrice EURIA, pour son suivi, ainsi que le temps qu'elle a su m'accorder notamment pour la relecture de mon mémoire.

Je remercie enfin ma famille et mes proches pour leur soutien dans cette période.

# Table des matières

Résumé	i
Abstract	iii
Note de synthèse	v
Summary note	xii
Remerciements	xviii
Table des matières	xxi
Introduction	1
<b>1 Contexte réglementaire</b>	<b>3</b>
1.1 Solvabilité 2	3
1.1.1 Généralités sur la norme Solvabilité 2	3
1.1.2 Zoom sur le pilier 1 : exigences quantitatives	4
1.2 Calcul du capital économique d'une compagnie d'assurance	9
1.2.1 Formule Standard	9
1.2.2 Modèle Interne	11
<b>2 Projection de la distribution du <i>Best Estimate</i> stressé à horizon 1 an</b>	<b>13</b>
2.1 Méthode des Simulations dans les Simulations	14
2.1.1 Premier niveau de simulation : en univers monde réel	15
2.1.2 Deuxième niveau de simulation : en univers risque neutre	15
2.1.3 Limites de cette méthode	15
2.2 Méthodes alternatives	16
2.2.1 Accélérateur de SdS	16
2.2.2 <i>Replicating Portfolio</i>	18
2.2.3 Approches paramétriques	19
2.3 Approche adoptée par Generali	21
2.3.1 Caractéristiques de la ligne d'activité étudiée	21
2.3.2 Identification des risques	21

2.3.3	Projection de la distribution du <i>Best Estimate</i> stressé par LSMC 2	27
2.3.4	Obtention de la distribution complète des Fonds Propres stressés	31
2.3.5	Déduction du capital de solvabilité requis (SCR)	32
<b>3</b>	<b>Utilisation de modèles de <i>Machine Learning</i> pour défier l'étape de régression du LSMC</b>	<b>35</b>
3.1	Données	36
3.1.1	Base d'apprentissage	36
3.1.2	Base de validation	38
3.1.3	Analyse univariée	38
3.2	Théorie des modèles utilisés	43
3.2.1	Forêt aléatoire	43
3.2.2	<i>EXtreme Gradient BOOSTing</i> (XGBoost)	47
3.2.3	<i>Multivariate Adaptive Regression Splines</i> (MARS)	50
3.2.4	<i>Support Vector Machine</i> : SVM	53
3.3	Calibration des modèles et résultats obtenus	56
3.3.1	Validation croisée par k-fold	56
3.3.2	Modèles finaux et résultats graphiques	57
3.4	Comparaison et choix du meilleur modèle	61
3.4.1	Métriques utilisées pour la comparaison	61
3.4.2	Résultats obtenus et comparaison des modèles	62
3.4.3	Choix et étude du meilleur modèle	64
<b>4</b>	<b>Impact d'un changement de forme paramétrique sur le SCR final</b>	<b>67</b>
4.1	Différence entre la fonction polynomiale actuelle et la fonction obtenue par MARS	67
4.1.1	Le polynôme actuel en détails	67
4.1.2	La fonction de régression calibrée par le méthode MARS en détails	68
4.1.3	Principales différences entre les deux fonctions répliquatives	69
4.2	Estimation du SCR avec la forme paramétrique obtenue par la régression MARS	72
4.2.1	Processus de calcul	72
4.2.2	Résultats obtenus	73
4.3	Impact du changement de forme paramétrique	73
4.3.1	Comparaison	74
4.3.2	Analyse de l'écart observé	76
<b>5</b>	<b>Conclusion</b>	<b>83</b>
5.1	Conclusion	83
5.2	Pour aller plus loin	85
<b>A</b>	<b>Fonctionnement d'un arbre de décision pour la régression</b>	<b>87</b>

<b>B Intervalles de <i>fitting</i> pour les modèles calibrés sur la ligne d'activité assurance vie</b>	<b>88</b>
<b>C Construction de la <i>Risk Factor Scenario Set</i> (RFSS)</b>	<b>89</b>
<b>D Résultats du modèle MARS sur les autres lignes d'activités</b>	<b>91</b>
D.1 Ligne d'activité regroupant les contrats de protection : LoB PRO . . . . .	91
D.2 Ligne d'activité regroupant les contrats de retraite : LoB PENUN . . . . .	92
D.3 Ligne d'activité regroupant les contrats de retraite professionnelle : LoB PENTR . . . . .	93
D.4 Ligne d'activité regroupant les contrats en unités de compte : LoB UC . .	95
<b>Glossaire</b>	<b>97</b>
<b>Bibliographie</b>	<b>102</b>





# Introduction

Le contexte réglementaire assurantiel et notamment la norme européenne Solvabilité 2 impose aux compagnies d'assurance de suivre les risques auxquels elles sont exposées en continu. En particulier, l'une des exigences quantitatives de Solvabilité 2 est le calcul du *SCR*, *Solvency Capital Requirement*, correspondant au montant de capital minimal que doit détenir un assureur pour que sa probabilité de ruine économique à horizon 1 an soit inférieure à 0.5%. Le calcul de ce besoin en capital peut s'effectuer par l'application de la Formule Standard Solvabilité 2, ou par une approche par modèle interne nécessitant l'approbation préalable des autorités de contrôle.

Generali utilise un modèle interne pour le calcul des exigences réglementaires. La valorisation du besoin en capital nécessite alors la projection de la distribution des fonds propres à 1 an. Or, compte tenu des interactions existantes entre les actifs et les passifs de l'assureur qui imposent une valorisation stochastique de ses engagements, le calcul du *SCR* implique la réalisation de simulations stochastiques imbriquées (un premier jeu de simulations projetant les facteurs de risque de  $t=0$  à  $t=1$  an en univers monde réel, puis pour chacun de ces scénarios un second jeu de simulation en univers risque neutre pour valoriser les postes du bilan prudentiel). La réalisation de ces deux niveaux de simulations imbriquées pose de vrais problèmes opérationnels à cause de la lourdeur et du temps computationnel.

Afin d'accélérer l'estimation de la distribution des passifs à 1 an, Generali utilise une approche paramétrique appelée *Least Square Monte Carlo* visant à réduire le nombre de simulations secondaires. Le but de cette méthode est d'obtenir une forme paramétrique du *Best Estimate* en fonction des facteurs de risque, afin de pouvoir le valoriser instantanément. La méthode consiste alors à ne faire que quelques simulations risque neutre pour chaque simulation primaire afin de calculer un *BE* approximatif pour chacun des scénarios de stress primaire. L'augmentation de l'erreur due au faible nombre de simulations stochastiques pour la valorisation est corrigée par une régression des moindres carrés sur ces valeurs imprécises, débouchant sur une fonction paramétrique permettant de calculer le *Best Estimate* stressé en fonction des différents facteurs de risques. À partir de cette fonction paramétrique il est possible de reconstruire la distribution complète des fonds propres et ainsi de calculer le SCR grâce au quantile à 0.5% de la distribution estimée des fonds propres.

Aujourd'hui, la fonction paramétrique est obtenue par une régression polynomiale

incluant des termes croisés, dont les termes sont choisis intelligemment par un algorithme de sélection de variables. La problématique de cette étude est de chercher à savoir si des algorithmes de *Machine Learning* peuvent optimiser la phase de régression de l'approche LSMC et permettre l'obtention d'une meilleure fonction proxy, évaluant le BE stressé en fonction des différents facteurs de risque. Nous nous intéresserons à un potentiel gain en terme de précision, ainsi qu'en terme de temps de calcul.

# Chapitre 1

## Contexte réglementaire

Avant d'aborder le calcul du besoin en capital d'une compagnie d'assurance, il est nécessaire de comprendre ce que représente ce besoin en capital ainsi que le contexte réglementaire dans lequel il doit être établi. Ce chapitre, inspiré du cours de Contexte Réglementaire dispensé à l'EURIA par M. HOROWITZ et Mme. BADOURALY-KASSIM [13], a pour but d'introduire la directive européenne Solvabilité 2.

### 1.1 Solvabilité 2

La norme Solvabilité 2 a pris effet en janvier 2016. Elle a la volonté d'être au plus proche des risques de l'assureur et de promouvoir une bonne gestion des risques dans la gouvernance de l'entreprise. Pour cela, elle responsabilise les assureurs en leur donnant plus de liberté, en contrepartie d'un renforcement du dialogue prudentiel. Une autre motivation de la réforme Solvabilité 2 est l'harmonisation du contrôle des groupes dans l'espace économique Européen, afin de faciliter la comparaison des différents groupes européens.

#### 1.1.1 Généralités sur la norme Solvabilité 2

Cette norme prudentielle est basée sur 3 piliers fondamentaux :

##### **Pilier 1 : Exigences quantitatives**

L'assureur doit évaluer ses actifs et ses passifs en valeur de marché dans son bilan prudentiel avec pour objectif de définir un niveau de fonds propres permettant de couvrir les exigences en capital réglementaires : le *SCR*, *Solvency Capital Requirement* et le *MCR*, *Minimum Capital Requirement*. Les provisions techniques doivent être valorisées de manière *Best Estimate*.

##### **Pilier 2 : Exigences qualitatives**

Plusieurs notions sont abordées dans ce pilier : le renforcement de la gouvernance avec notamment la définition de quatre fonctions clés indispensables dans la compagnie (Gestion des risques, Conformité, Audit interne et Actuariat), le principe

des 4 yeux (imposant 2 dirigeants effectifs), le renforcement du contrôle interne et de la gestion des risques de la compagnie ainsi qu'une auto-évaluation des besoins en capital (*Own Risk and Solvency Assessment*, ORSA). Enfin, l'application du principe de la personne prudente à la place de la limitation d'actif.

### **Pilier 3 : *Reporting* prudentiel et information du public**

Ce dernier pilier vise à renforcer la transparence aux yeux des contrôleurs ainsi que du public, par l'intermédiaire de plusieurs rapports sur la solvabilité et l'état financier de l'assureur à partager aux autorités de contrôle et/ou au public en fonction des rapports.

Le contexte réglementaire régissant la suite de ce mémoire étant concentré dans le pilier 1, les piliers 2 et 3 ne seront pas détaillés ici. Le lecteur intéressé pourra se référer à la Directive 2009/138/CE du Parlement européen et du Conseil du 25 novembre 2009 [7] pour plus d'informations.

## **1.1.2 Zoom sur le pilier 1 : exigences quantitatives**

### **Le bilan prudentiel**

Le bilan Solvabilité 2 est représenté de manière simplifiée en figure 1.1 ci-dessous.

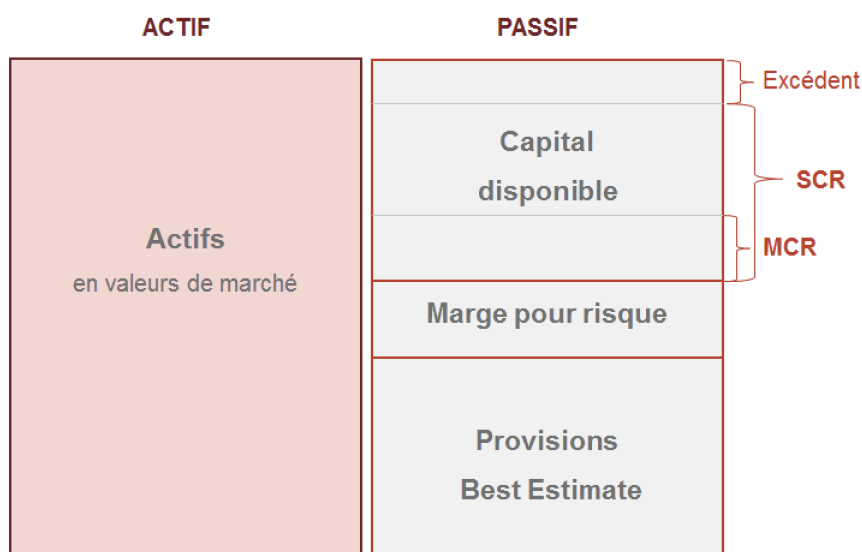


FIGURE 1.1 – Bilan prudentiel simplifié

Contrairement au bilan comptable, le bilan prudentiel comptabilise les actifs en valeurs de marché. Quant à elles, les provisions techniques sous Solvabilité 2 sont la somme de la meilleure estimation des engagements de l'assureur ("*Best Estimate*") et d'une

marge pour risque ("*Risk Margin*").

### ***Les exigences de capital : SCR et MCR***

Comme il a été dit lors de l'introduction à chapitre, outre imposer le cadre de calcul des différents postes du bilan prudentiel, le pilier 1 de la directive solvabilité 2 impose deux exigences réglementaires sur le capital économique détenu par la compagnie : le SCR et le MCR.

Le MCR (*Minimum Capital Requirement*), correspond au minimum absolu de capital à détenir. Le MCR est calculé à partir de formules distinguant les activités vie et non vie. Il ne peut être inférieur à 25% du SCR sans excéder 45% du SCR, et a un montant plancher minimum.

Le non respect de la couverture du MCR conduit à un retrait de l'agrément d'assureur par l'ACPR au bout de 3 mois.<sup>1</sup>

Le SCR (*Solvency Capital Requirement*), correspond à la *value at risk* (VaR) à horizon 1 an des fonds propres de base de l'entreprise avec un niveau de confiance de 99.5%. La VaR au seuil  $\alpha$  peut être vue comme le quantile  $1 - \alpha$  de la distribution des Fonds Propres de la compagnie. Autrement dit, il s'agit du montant de capital minimum dont l'assureur doit disposer pour que sa probabilité de ruine à horizon 1 an soit inférieure ou égale à 0.5%.

### ***Les exigences de capital : le ratio de solvabilité***

Le ratio de solvabilité est un indicateur clef de la solvabilité d'une compagnie d'assurance, il est exprimé en pourcentage et traduit de la santé financière de l'assureur. Également appelé ratio de couverture du SCR, il est obtenu en faisant le rapport des Fonds Propres de l'assureur sur son SCR.

$$\text{Ratio de solvabilité} = \frac{\text{Fonds Propres économiques éligibles}}{\text{SCR}}$$

Le montant minimum de ce ratio est de 100%, faute de quoi la compagnie est déclarée non solvable.

### ***La marge pour risque ou Risk Margin***

La marge pour risque correspond à la part des provisions techniques représentative du coût en capital qu'engendre le portage des risques, étant donné la réglementation. L'idée est qu'un éventuel organisme reprenneur des engagements devrait disposer de fonds propres, immobilisés à un coût supérieur aux produits qu'elle en tirerait.

### ***La provision Best Estimate***

La provision *Best Estimate*, communément appelée BE, est le poste du bilan sur lequel sera concentrée la suite de cette étude. Il correspond à « *la moyenne pondérée*

---

1. Source : Art. L. 325-1 Code des assurances

par leur probabilité des flux de trésorerie futurs compte tenu de la valeur temporelle de l'argent estimée sur la base de la courbe des taux sans risque pertinente, soit la valeur actuelle attendue des flux de trésorerie futurs »<sup>2</sup>.

Pour calculer le BE, les entreprises doivent segmenter leurs engagements en groupes homogènes de risques, au minimum par lignes d'activité (LoB : Line of Business). Les flux du *Best Estimate* doivent tenir compte de l'ensemble des flux de trésorerie futurs des contrats d'assurance et de réassurance existants :

- (-) les primes
- (+) les prestations (règlement de sinistres, rentes servies, etc)
- (+) les frais (administratifs et de gestion).

En assurance vie, le calcul du BE doit également tenir compte des options contractuelles et des garanties financières du contrat, comme la participation aux bénéfices, le taux minimum garanti, ou encore la possibilité de rachat anticipé du contrat. La présence de ces options et garanties introduit de fortes interactions entre le passif et l'actif de l'assureur. En effet, le code des assurances impose aux assureurs de reverser au moins 85% des bénéfices financiers (issus des placements effectués avec les fonds des assurés) et au moins 90% des bénéfices techniques aux assurés. De ce fait, en cas de rendement financier favorable, l'assureur devra reverser une participation aux bénéfices aux assurés, à hauteur minimum de 85%, ce qui va diminuer les gains de l'assureur. Alors que dans le cas d'un rendement défavorable, l'assureur a l'obligation de respecter ses engagements en revalorisant les contrats au taux minimum garanti (TMG) stipulé dans le contrat.

Du fait de ces interactions, créant un déséquilibre des engagements de l'assureur envers les assurés, le calcul des provisions *Best Estimate* doit être fait de manière stochastique afin de tenir compte du coût des options et garanties.

D'un point de vue mathématiques, il est calculé par la formule suivante :

$$BE_t = E^Q \left[ \sum_{i=t+1} \frac{CF_i \times p_i}{(1 + \tau_i)^i} \mid Y_t \right]$$

avec

- $CF_i$  le Cash Flow (flux de trésorerie) sortant en  $i$  avec la probabilité  $p_i$
- $Q$  la probabilité risque neutre,
- $Y_t$  les risques financiers en  $t$ ,
- $\tau_i$  le taux sans risque pertinent au temps  $i$ .

---

2. Source : Art. R351-2 Code des assurances

D'un point de vue opérationnel, le *Best Estimate* est valorisé par son estimateur Monte Carlo, puisque cette quantité ne peut être valorisé de manière exacte par formule fermée compte tenu de la complexité des interactions entre le passif et l'actif. L'approche Monte Carlo consiste à effectuer un grand nombre de simulations stochastiques pour approcher une espérance. En effet, comme le démontre la loi des Grands Nombres, sous certaines hypothèses l'espérance d'une variable aléatoire peut être estimée par la moyenne sur un très grand nombre de réalisations indépendantes et identiquement distribuées de cette variable aléatoire.

Pour ce faire, l'utilisation d'un générateur de scénarios économiques (GSE) est requise afin de simuler N réalisations indépendantes de prix des actifs financiers à considérer sous la probabilité risque neutre. Le générateur de scénarios économiques va produire un prix simulé sur une période T pour chaque scénario et pour chaque actif.

La valorisation « Risque Neutre » consiste à simuler les flux futurs de manière « market consistent » et à considérer la moyenne des flux futurs actualisés au taux sans risque. Dans un univers risque neutre, les hypothèses d'absence d'opportunité d'arbitrage (AOA<sup>3</sup>) et d'efficacité du marché sont faites, dans ces conditions les prix futurs actualisés au taux sans risque sont alors martingales.

Les scénarios économiques générés par le GSE sont ensuite introduits dans le modèle de projection ALM de l'entreprise afin de simuler en particulier les flux de trésorerie futurs dans chacun des scénarios, sur un horizon temporel donné (30 ans dans le cadre du modèle ALM de Generali).

L'estimateur de Monte Carlo du *Best Estimate* est obtenu par la formule suivante :

$$\hat{BE}_t = \frac{1}{N} \sum_{k=1}^N \left[ \sum_{i=t+1}^Y \frac{CF_{i,Y}}{(1 + \tau_i)^i} \right] \quad (1.1)$$

avec

- $CF_{i,Y_i,n}$  les flux probables sortant en i conditionnellement au scénario économique k en i
- N le nombre de simulations risque neutre

### Accélérateur de convergence : utilisation de variables antithétiques

L'utilisation de variables antithétiques est une méthode de réduction de variance de l'estimateur de Monte-Carlo. L'idée est de tirer parti de certaines symétries de distribution et de corrélations négatives entre deux variables aléatoires, notamment de la propriété de symétrie du mouvement brownien  $W =^{loi} -W$ .

Le principe est le suivant : supposons que l'on dispose d'une variable aléatoire U suivant une loi uniforme sur [0,1], et que nous cherchions à calculer  $I = E[f(U)]$ . Pour

---

3. AOA : Il est impossible d'obtenir un profit futur certain pour une mise initiale nulle

calculer cette espérance par son estimateur de Monte-Carlo, nous allons calculer :

$$\hat{I} = \frac{1}{N} \sum_{i=1}^N f(U_i)$$

Or, puisque  $1-U$  suit également une loi uniforme sur  $[0,1]$ , cela revient à calculer :

$$\hat{I}' = \frac{1}{N} \sum_{i=1}^N f(1 - U_i)$$

Ou encore, puisque  $E[f(U)] = E[f(1 - U)]$ ,  $I = \frac{1}{2}(E[f(U)] + E[f(1 - U)])$  peut être approché par :

$$\tilde{I} = \frac{1}{N} \sum_{i=1}^{[N/2]} (f(U_i) + f(1 - U_i))$$

En s'intéressant à la variance de l'estimateur  $\tilde{I}$ ,

$$\text{var}(\tilde{I}) = \text{var}(\hat{I}) + \frac{1}{N} \text{Cov}(f(U), f(1 - U))$$

Il est remarquable que  $\text{var}(\tilde{I}) \leq \text{var}(\hat{I})$  si et seulement si  $\text{Cov}(f(U), f(1 - U)) \leq 0$ . Or, si  $f$  est monotone alors  $\text{Cov}(f(U), f(1 - U)) \leq 0$ . Et cette monotonie est en général observée dans le cadre de l'estimation *Best Estimate* des engagements d'un assureur vie. En effet, plus le prix de l'actif est élevé et plus l'engagement de l'assureur va généralement être élevé dû aux participations aux bénéfices financiers ou dû à un taux d'actualisation plus faible.

Il ressort de cela que l'utilisation de variables antithétiques permet de diminuer la variance de l'estimateur de Monte-Carlo classique, tout en simulant deux fois moins de variables aléatoires. De ce fait cette méthode de réduction de variance est une pratique très répandue au sein des assureurs pour l'évaluation de leur passifs.

Considérons un scénario économique produit à partir de nombres pseudo-aléatoires qui sont générés en tirant des réalisations  $(u_i)_{i \in [1, N]}$  d'une variable uniforme sur  $[0,1]$  et en leur appliquant l'inverse des fonctions de répartition des lois que l'on souhaite simuler :

$$\gamma_i = F^{-1}(u_i)$$

Un scénario économique est alors une fonction de ces réalisations, représentant les prix des actifs financiers au cours du temps :

$$(x_i^t) = f(\gamma_1, \dots, \gamma_p)$$

Le scénario antithétique correspondant à ce premier scénario pourra être construit en réutilisant les mêmes réalisations  $(u_i)_{i \in [1, N]}$  de la loi uniforme et en déterminant les grandeurs :

$$\gamma'_i = F^{-1}(1 - u_i)$$

et les prix des actifs dans le temps :

$$(x_i'^t) = f(\gamma'_1, \dots, \gamma'_p)$$



## 1.2 Calcul du capital économique d'une compagnie d'assurance

La section suivante a pour objectif d'expliquer les procédés de calcul du SCR, ses fondements reposent sur le cours de Modélisation et agrégation des risques [4], dispensé par M. BENNOUNNA et L. DEVINEAU à l'EURIA.

Comme nous l'avons vu dans la section précédente, le capital économique requis de Solvabilité 2 correspond au montant de fonds propres que doit détenir un assureur pour faire face à une ruine économique bicentenaire, à horizon 1 an. La ruine économique survient lorsque la valeur de marché des actifs est inférieure à la valeur des engagements *Best Estimate* du passif.

Mathématiquement, le capital économique peut être formalisé par la formule suivante :

$$C = FP_0 - q_{0.5\%}(D(1) \times FP_1) \quad (1.2)$$

- $FP_0$  les Fonds Propres économiques en  $t = 0$ ,
- $FP_1$  les Fonds Propres économiques en  $t = 1$ ,
- $D(1)$  le facteur d'actualisation de première période,
- $q_{0.5\%}$  le quantile au seuil 0.5%.

Pour estimer la valeur du capital économique, deux approches peuvent être envisagées par l'assureur : le calcul par Formule Standard ou le calcul par Modèle Interne, spécifique à l'entreprise. L'approche par Formule Standard sera abordée sans entrer dans les détails, car l'approche suivie dans ce mémoire sera celle par Modèle Interne.

Certaines hypothèses de calcul sont communes aux deux méthodes : l'horizon temporelle à 1 an, la continuité d'exploitation, la prise en compte de l'effet de diversification des risques due aux corrélations entre les risques (la probabilité que deux risques décorrélés subissent une dégradation bicentenaire en même temps est bien inférieure à 0.5%). De plus, le SCR a l'obligation d'être suivi régulièrement et doit être immédiatement recalculé en cas de déviation du profil de risque de l'entreprise.

### 1.2.1 Formule Standard

La Formule Standard est une méthode de calcul qui est commune à chaque compagnie d'assurance. Cette approche contourne la projection du bilan économique en calculant le capital requis via l'application de différents chocs bicentenaires au *Best Estimate en  $t=0^+$* . Le niveau de capital requis est ensuite obtenu en faisant la différence entre le BE central et le BE choqué.

C'est une approche modulaire qui repose sur trois étapes fondamentales :

- **Première étape** : calcul d'un capital économique pour chaque *risque élémentaire*.<sup>4</sup>
- **Deuxième étape** : Agrégation intra-modulaire, les différents SCR élémentaires sont agrégés au sein de chaque module de risque à l'aide de matrices de corrélations calibrées par l'EIOPA.
- **Troisième étape** : Agrégation inter-modulaire, les SCR de chacun des modules de risques sont agrégés entre eux à l'aide d'une matrice de corrélation spécifiée dans le règlement délégué, afin de former le BSCR : *Basic Solvency Capital Requirement*.

⇒ Obtention du besoin en capital global par la formule :

$$SCR = BSCR + SCR_{op} - Adj$$

où  $SCR_{op}$  représente le besoin en capital pour couvrir le risque opérationnel, et  $Adj$  est l'ajustement au titre de l'absorption des pertes par impôts différés et par participation aux bénéfices.

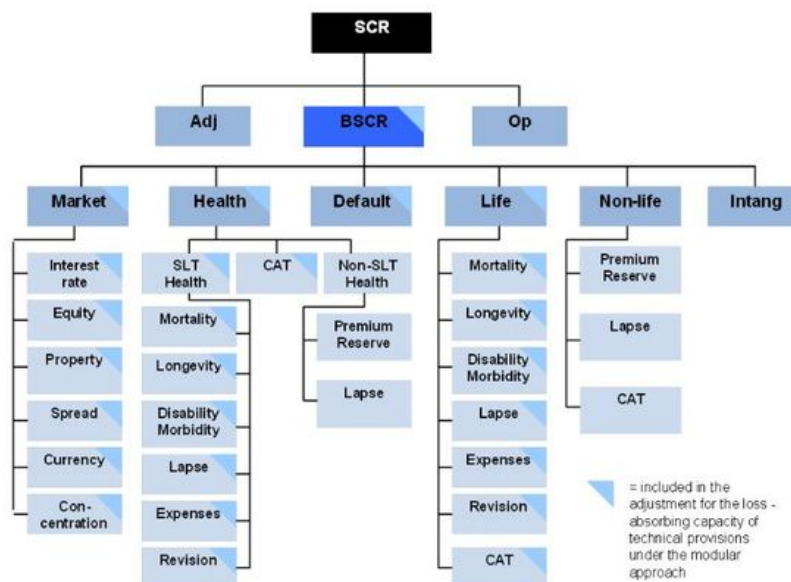


FIGURE 1.2 – Agrégation des risques en formule standard

Les différents SCR élémentaires représentent le besoin en capital au titre d'un seul risque. Ils sont calculés par la différence entre une NAV (*Net Asset Value*) choquée sur ce facteur de risque et la NAV centrale. Les chocs à appliquer sont imposés par le régulateur et détaillés dans le Règlement délégué (UE) 2015/35 de la commission du 10 octobre 2014 [2], ils représentent des chocs bicentennaires. Pour plus de détails sur le calcul du capital

4. Les risques élémentaires sont ceux représentés en bas de la *pieuvre* présentée en figure 1.2

## 1.2. CALCUL DU CAPITAL ÉCONOMIQUE D'UNE COMPAGNIE D'ASSURANCE

---

économique par la formule standard, le lecteur pourra se référer au règlement délégué en question.

### 1.2.2 Modèle Interne

Lorsque l'assureur juge que la formule standard n'est pas correctement adaptée à son profil de risque, il peut choisir de calculer les exigences réglementaires par un modèle interne, partiel ou total. Ce modèle interne, qu'il soit partiel ou total, doit être approuvé par l'autorité de contrôle préalablement à son utilisation.

Bien que le calcul du capital économique par un modèle interne soit plus proche du risque spécifique de la compagnie, sa mise en place est complexe et très coûteuse, ce qui explique l'emploi de la formule standard chez un grand nombre d'assureurs.

#### *Approche modulaire vs approche par projection de la distribution de fonds propres à 1 an*

Deux types de méthodes sont généralement envisagées dans le calcul du capital économique en modèle interne :

- Soit une approche modulaire, suivant le même principe que la formule standard, avec des risques élémentaires et des matrices de corrélations spécifiques à l'entreprise ;
- Soit une méthode par projection de la distribution des fonds propres économiques à horizon 1 an, afin d'en calculer le quantile à 0.5% et ainsi d'en déduire le besoin en capital (SCR). Contrairement à une approche modulaire, la projection de la distribution des fonds propres est obtenue en faisant varier les risques simultanément.

Le chapitre suivant vise à entrer dans le détail du calcul du besoin en capital en vision modèle interne suivant la méthode par projection de la distribution des fonds propres à horizon 1 an. Plus précisément, ce chapitre décrira les différentes méthodes utilisées par les assureurs pour projeter la distribution de la provision *Best Estimate* du passif à un an. La distribution des fonds propres à 1 an sera alors déduite de celle-ci, après valorisation des actifs à 1 an par soustraction du passif à l'actif :

$$FP_{t=1} = VM\_Actif_{t=1} - (BE_{t=1} + RM_{t=1})$$

avec FP les fonds propres, VM\_Actif la valeur de marché des actifs, BE le Best Estimate et RM la marge pour risque.

**CONCLUSION DU CHAPITRE :**

La connaissance du contexte réglementaire et notamment des exigences de la directive Solvabilité 2 est essentielle pour comprendre où s'inscrivent les recherches effectuées dans ce mémoire. En effet, Solvabilité 2 impose aux assureurs européens un suivi constant de leurs risques et de leur solvabilité. Entre autre, la directive demande de calculer un montant de capital minimum à détenir pour que la probabilité de ruine économique à un horizon de 1 an soit inférieure ou égale à 0.5% : le SCR. De plus, ce montant de SCR doit être suivi tout au long de l'année. Dans le cadre d'un calcul par modèle interne, ce qui est le cas de Generali, la valeur de SCR s'obtient par projection de la distribution des Fonds Propres, ou du passif *Best Estimate*, à 1 an.

**TRANSITION :**

*Du fait de l'existence d'interactions entre le passif et l'actif d'un assureur vie, l'obtention de la distribution complète des Fonds Propres, ou du Best Estimate, est complexe. Le chapitre suivant vise à expliquer les différents procédés permettant de projeter la distribution du passif Best Estimate à 1 an dans le but final de calculer le SCR de l'entreprise. L'approche utilisée par Generali sera détaillée en fin de chapitre.*

## Chapitre 2

# Projection de la distribution du *Best Estimate* stressé à horizon 1 an

Les options et garanties incluses dans les contrats d'assurance vie telles que la participation aux bénéficiaires, le taux minimum garanti ou encore la possibilité de rachat anticipé, créent de nombreuses interactions entre l'actif et le passif de l'assureur. L'existence de ces interactions imposent de valoriser le bilan assurantiel de manière stochastique afin de prendre en compte la valeur temps des options et garanties, notamment par la réalisation de simulations de Monte Carlo comme il a été dit dans la section 1.1.2.

De plus, la valorisation des postes du bilan à  $t=1$  nécessite de connaître l'environnement économique à 1 an. Des scénarios économiques sont alors générés par un GSE (générateur de scénarios économiques) afin d'accéder aux valeurs des facteurs de risques à 1 an.

Cela amène à effectuer des simulations stochastiques imbriquées (*Nested Simulations*) afin d'obtenir la distribution de la provision *Best Estimate* du passif de l'assureur, cette méthode appelée Simulations dans les Simulations ou SdS sera présentée en section 1. La réalisation de ces deux jeux de simulations imbriquées entraîne une telle lourdeur de calcul que le recours à des méthodes d'approximation permettant de se substituer à la méthode des SdS est fréquent dans les compagnies. Quelques unes de ces méthodes seront expliquées dans les sections suivantes. La dernière section présentera plus en détail l'approche utilisée dans le modèle interne de Generali, pour projeter la distribution des passifs *Best Estimate* stressés.

### **NB : Application des chocs au temps $t=0^+$**

Un raccourci est fait dans le modèle interne, consistant à appliquer les chocs sur les facteurs de risque en  $t = 0^+$  et non pas en  $t = 1$ . Il n'y a donc pas de projection monde réel de l'environnement économique jusqu'à  $t=1$  an. Les scénarios primaires consistent seulement à appliquer des chocs sur les différents facteurs de risque en  $t = 0^+$ .

Cette approximation est fréquente chez les acteurs de l'assurance utilisant un modèle interne, et c'est notamment le cas chez Generali. Les impacts engendrés par cette simplification sont supposés négligeables.

Dans le suite de ce mémoire, nous continuerons de parler du *Best Estimate* et des Fonds Propres stressés à 1 an afin de s'aligner à la littérature et à la réglementation Solvabilité 2, mais il faudra garder en tête que de manière opérationnelle il s'agit du temps  $t = 0^+$  dans le modèle interne de Generali.

## 2.1 Méthode des Simulations dans les Simulations

La méthode des Simulations dans les Simulations, notée SdS, fonctionne par deux niveaux de simulations imbriqués. Le premier niveau de simulations, appelé simulations primaires, a pour but de projeter l'univers économique à  $t=1$  an. Ce premier jeu de simulations est effectué en univers Monde Réel, ce qui signifie qu'elles sont construites à partir de l'observation de l'historique des indices des différentes variables économiques, afin d'en donner une évolution réaliste. Une fois ce premier niveau de simulations effectué, il est nécessaire de valoriser le bilan prudentiel de l'assureur en  $t=1$  an et en particulier la valeur *Best Estimate* du passif, dans chacun des scénarios résultant des simulations primaires. Pour valoriser le passif de l'assureur, il est alors nécessaire de réaliser des simulations de Monte Carlo conditionnellement à l'environnement économique de chaque scénario primaire : elles correspondent au deuxième niveau de simulation, les simulations secondaires, réalisées dans un univers Risque Neutre. La projection risque neutre part du principe que les investisseurs sont neutres au risque dans le choix de leurs investissements.

La figure suivante, extraite du cours de modélisation et d'agrégation des risque [4], schématise le fonctionnement de l'approche des Simulations dans les Simulations.

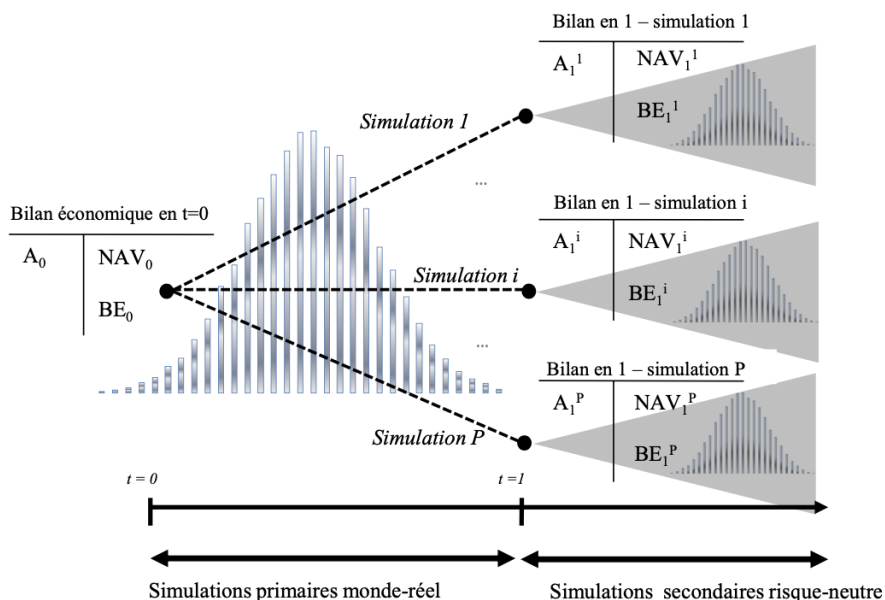


FIGURE 2.1 – Schématisation de l'approche des Simulations dans les Simulations

### 2.1.1 Premier niveau de simulation : en univers monde réel

Afin de calculer le *Best Estimate* à 1 an, il est nécessaire dans un premier temps de connaître les conditions économiques à ce moment. C'est pourquoi la première étape consiste à projeter l'environnement par  $N$  simulations, correspondant à des scénarios économiques différents. Ces simulations sont appelées les *simulations primaires* et sont effectuées en monde réel, c'est à dire à partir de l'historique ou à dire d'expert, afin d'être au plus proche de la réalité comme l'exige la réglementation.

### 2.1.2 Deuxième niveau de simulation : en univers risque neutre

Une fois que l'environnement économique est projeté, il faut reconstituer le bilan prudentiel à un an pour chacun des  $N$  scénarios primaires. Pour cela, il faut notamment re-calibrer un générateur de scénarios économique risque neutre dans l'optique de projeter les scénarios risque neutre sur un horizon temporel donné, afin de réévaluer les passifs d'assurance vie avec options et garanties après l'application d'un choc en monde réel. Pour valoriser le passif de l'assureur et notamment la provision *Best Estimate* de manière précise, la méthode des SdS procède par la réalisation d'un jeu complet de  $M$  simulations de Monte Carlo. Ces simulations Monte-Carlo correspondent alors au deuxième niveau de simulation.

### 2.1.3 Limites de cette méthode

Cette méthode, bien que la plus naturelle et la plus exacte mathématiquement, possède une limite opérationnelle : le temps de calcul. En effet, les deux niveaux de simulations imbriqués engendrent une complexité opérationnelle considérable, nécessitant un budget de  $N \times M$  simulations. Sachant que la convergence des simulations Monte Carlo nécessite généralement un minimum de  $M = 1\,000$  à  $5\,000$  simulations<sup>1</sup>, et le jeu de simulations primaires contient la plupart du temps entre  $10\,000$  et  $50\,000$  simulations pour être représentatif. L'utilisation de l'approche des Simulations dans les Simulations pour le calcul du capital économique requis (SCR) demanderait alors plus de  $10\,000\,000$  de simulations.

Il est important de comprendre qu'au cours d'une simulation de nombreux calculs sont opérés au sein du modèle ALM de l'entreprise. En effet, le modèle ALM a pour objectif de projeter sur un horizon temporel défini (30 ans ici) tous les flux de trésorerie de la compagnie d'assurance. Pour cela, il est nécessaire de modéliser un grand nombre d'hypothèses sur l'actif, sur le passif ainsi que sur la corrélation entre les deux. Coté actif, cela implique la projection de l'environnement économique sur 30 ans, ou encore les règles de réallocation d'actifs. Coté passif, les prestations futures sont modélisées par l'intermédiaire, entre autre, de table de mortalité. Le comportement des assurés est lui aussi modélisé à partir de loi de rachat tenant compte de la situation économique, et

---

1. Le nombre  $M$  de simulations Monte Carlo nécessaires dépend de la constitution du portefeuille d'actif. Des tests sont exigés annuellement par l'ACPR pour challenger la valeur de  $M$ .

notamment du taux de revalorisation des contrats. Cette complexité implique un temps de calcul conséquent, ce qui explique pourquoi la réalisation d'autant de simulations dans la méthode des SdS est problématique.

Pour un portefeuille d'assurance vie comprenant des interactions complexes entre l'actif et le passif, un tel budget de simulations est irréalisable d'un point de vue opérationnel. Pour cette raison, il est intéressant voire obligatoire de s'intéresser aux méthodes alternatives.

## 2.2 Méthodes alternatives

Afin de pallier au problème de temps de calcul inhérent à la méthode des Simulations dans les Simulations, plusieurs méthodes alternatives permettant de valoriser le montant de capital de solvabilité requis (SCR) ont vu le jour dans la littérature actuarielle. Plusieurs de ces méthodes alternatives seront présentées dans cette section. Puis l'une d'entre elle, la méthode de *Least Square Monte Carlo* (LSMC), sera détaillée étant l'approche adoptée au cours de ce mémoire.

### 2.2.1 Accélérateur de SdS

L'accélérateur de SdS a pour but d'accélérer l'approche des Simulations dans les Simulations, et cela en localisant de façon suffisamment précise le scénario qui conduit au quantile de la distribution des Fonds Propres en  $t=1$  avec le moins de calculs ALM possible. Tous les axes d'optimisation de gain de temps sont explorés :

- Réduction du nombre de simulations primaires : les situations extrêmes sont identifiées a priori (queue de distribution). Les simulations centrales ne sont pas jouées dans le modèle ALM stochastique.
- Réduction du nombre de simulations secondaires : calcul avec un nombre de simulations assurant la *Market Consistency* (milliers de simulations) uniquement au voisinage du quantile, pour les autres points, utilisation d'un nombre réduit de simulations secondaires.
- Réduction du temps de calibration des tables secondaires : les scénarios secondaires sont obtenus par ajustement d'une unique table risque neutre calibrée à  $t=0$  et comportant un nombre réduit de simulations.

Cette approche, élaborée par L. DEVINEAU et S. LOISEL [5], procède par les étapes suivantes, en partant d'un ensemble de 15 000 scénarios monde réel :

- **Localisation du scénario primaire quantile :**

La première étape de l'approche d'accélérateur de SdS consiste à localiser les scénarios amenant à des valeurs de passifs extrêmes. Pour cela, seulement 50 scénarios



secondaires risque neutre sont effectués pour chacune des 15 000 simulations primaires. Par ailleurs, les scénarios secondaires sont obtenus par ajustement d'une unique table risque neutre calibrée à  $t=0$  et comportant un nombre réduit de simulations. Le nombre de simulations primaires est également diminué en procédant par les itérations suivantes, qu'il est possible de visualiser grâce au schéma figure 2.2 ci-dessous :

1. (points bleus sur le schéma) Simulations SdS sur le jeu des 300 scénarios primaires les plus adverses<sup>2</sup> sur 50 simulations secondaires avec méthode des ajustements.
2. (points rouges sur le schéma) Simulation des 300 simulations adverses suivantes. Si les 75<sup>3</sup> pires valeurs des 600 simulations effectuées sont les mêmes que celles des 300 premiers (itération précédente), arrêt de l'algorithme : scénario quantile estimé = couple de la 75ème pire valeur.  
Sinon :
3. (points vert sur le schéma) Répétition du processus jusqu'à ce que les 75 pires valeurs de l'itération correspondent aux 75 pires valeurs de l'itération précédente.

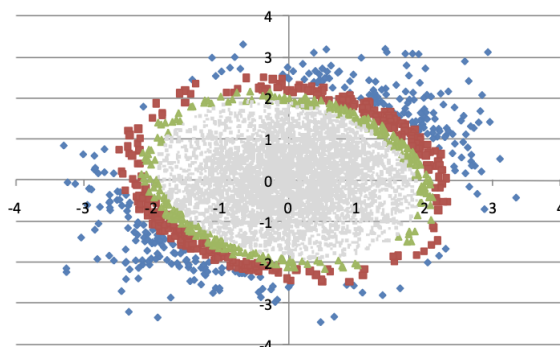


FIGURE 2.2 – Schématisation de la localisation du scénario primaire quantile  
Source : Cours de modélisation et d'agrégation des risques, M2 EURIA

#### • Projection SdS sur tables secondaires re-générées :

Grâce aux résultats de l'étape précédente, on dispose de la queue de distribution des Fonds Propres dans laquelle se trouve le quantile à 0.5%. Les Fonds Propres de l'étape précédente ont été calculés de manière approchée, à l'aide de tables

---

2. Afin de classer les scénarios, une norme est définie (basée sur les valeurs des aléas taux et actions, ainsi que sur la sensibilité des Fonds Propres à chacun de ces risques).

Les scénarios de norme élevée correspondent aux points en périphérie du nuage.

3. La 75<sup>ème</sup> pire valeur correspond au quantile à 0.5% de la distribution des passifs sur les 15 000 scénarios primaires :  $0.5\% \times 15000 = 75$

secondaires comportant peu de simulations, qui sont ajustées et non pas re-calibrées en fonction des conditions économiques en date  $t=1$ .

La deuxième étape revient donc à se concentrer seulement sur la zone proche du quantile, afin de faire un calcul exact des Fonds Propres pour les points étant au voisinage du quantile. C'est à dire avec un nombre de simulations secondaires respectant la *Market Consistency* et une prise en compte de tables secondaires re-calibrées et non pas ajustées.

Le voisinage est défini de manière symétrique en considérant  $N$  points avant et  $N$  points après le quantile, afin que le quantile corresponde à la médiane du sous-échantillon constitué par les  $2N+1$  points du voisinage.

- **Détermination du capital économique :**

Lors de cette dernière étape, un classement de la distribution obtenue après le SdS re-généré sur les  $2N+1$  points sélectionnés est fait. La valeur du scénario quantile utilisée pour le calcul du capital économique est la valeur médiane du sous-échantillon pour lequel un calcul exact a été effectué.

### 2.2.2 *Replicating Portfolio*

L'objectif de la méthode de *Replicating Portfolio*, ou portefeuille répliquant, est de déterminer un portefeuille d'actifs financiers permettant de répliquer la valeur des Fonds Propres ou la valeur économique des passifs en  $t=1$ .

Un « portefeuille répliquant » est donc un portefeuille d'actifs qui reproduit les *cash flows* de passif pour chaque simulation et chaque date future. Ainsi, à l'horizon de ruine ( $t=1$ ), l'équation suivante est vérifiée :

$$\text{Valeur du portefeuille répliquant} = \text{Best Estimate des Passifs}$$

Le principe est alors d'utiliser ce portefeuille répliquant comme proxy des Fonds Propres ou des passifs afin de calculer le capital économique directement en évaluant la valeur de marché du portefeuille répliquant. Le recourt aux portefeuilles répliquants permet de projeter aisément le bilan puisque la valeur de celui-ci peut être déterminée rapidement à chaque date future grâce à l'utilisation de formules fermées (ou des techniques numériques peu consommatrices en temps de calcul), ce qui évite de projeter les passifs à l'aide de simulations risque-neutre pour évaluer le BE.

Afin de calibrer le portefeuille répliquant, il faut tout d'abord déterminer les actifs candidats (call, put, swaption ...) ainsi que leurs paramètres. Ensuite, l'*asset mix* du portefeuille répliquant est calculé par une régression des fonds propres (ou du *Best Estimate*) à  $t=1$  sur les actifs candidats. A ce stade, il est possible d'accéder à la distribution des Fonds Propres (ou du *Best Estimate*) à  $t=1$ , puis à la valeur du SCR à partir du quantile à 0.5% des Fonds Propres.

La limite des portefeuilles répliquants est qu'ils ne couvrent que les risques financiers et de crédit (*spread*) et pas les autres risques non-financiers (rachat, mortalité, etc ...). Il est également souvent difficile de faire correspondre correctement des passifs complexes avec des actifs répliquants, puisque certains instruments requis ne sont pas disponibles sur le marché. De plus, en vertu des règles de Solvabilité 2, cela devient encore plus difficile en raison des taux d'actualisation différents utilisés pour les actifs et les passifs.

### 2.2.3 Approches paramétriques

Les méthodes suivantes visent à obtenir la distribution des Fonds Propres ou du *Best Estimate* à l'aide d'une forme paramétrique (formule fermée). Deux approches paramétriques pour la projection du passif assurantiel vont être présentées, ces méthodes vont accélérer l'approche des Simulations dans les Simulations de deux manières différentes : la première cherchera à diminuer le nombre de simulations primaires et la forme paramétrique sera calibrée sur des valeurs de Fonds Propres économiques ou de *Best Estimate*, tandis que la seconde vise à réduire le nombre de simulations secondaires et la forme paramétrique sera alors calibrée sur des Valeurs Actuelles Net (VAN) de marges.

#### *Curve Fitting*

L'approche de *Curve Fitting* va chercher à calibrer une forme paramétrique permettant d'exprimer les Fonds Propres économiques en date  $t=1$  en fonction des facteurs de risques, sur un nombre restreint de valeurs de Fonds Propres.

Pour cela, un nombre restreint de simulations primaires extrêmes (une centaine) va être déterminé, puis, après un recalibrage de tables secondaires risque neutre, les valeurs de *Best Estimate* (ou de Fonds Propres) associées à ces scénarios sont calculées en utilisant la méthode de Monte Carlo avec le nombre usuel de simulations (quelques milliers). Une interpolation est ensuite calibrée à partir des valeurs de *Best Estimate* (ou de Fonds Propres) ainsi obtenues afin de construire une fonction paramétrique du *Best Estimate* stressé en fonction des différents facteurs de risques.

La figure suivante schématise le fonctionnement de l'approche *Curve Fitting*.

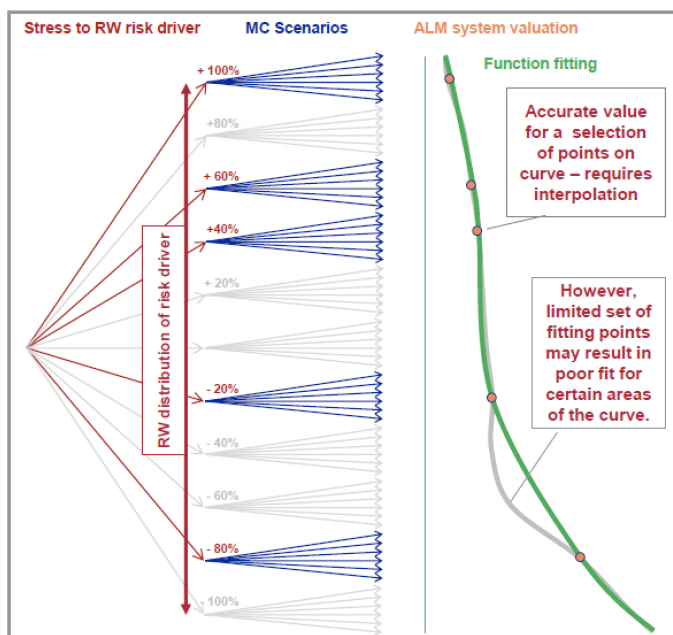


FIGURE 2.3 – Schématisation de l'approche *Curve Fitting*.

Source : Documentation interne à Generali sur la méthodologie du modèle interne

La limite de l'approche *Curve Fitting* est le choix des scénarios primaires extrêmes réalisés, reposant souvent sur des jugements d'experts. De plus, le faible nombre de points de *fitting* peut amener à une mauvaise fonction proxy dans certaines zones peu représentées dans les scénarios primaires. Cette seconde limite peut être visualisée sur le schéma de droite ci-dessus.

### *Least Square Monte Carlo*

Contrairement à l'approche *Curve Fitting*, l'approche de *Least Square Monte Carlo* (LSMC) vise à diminuer le nombre de simulations secondaires. La méthode consiste à effectuer un grand nombre de simulations primaires en monde réel, puis pour chacune de ces simulations, seulement quelques (souvent deux) simulations secondaires en risque neutre. Cela va permettre d'obtenir pour chaque simulation primaire une valeur de *Best Estimate* inexacte<sup>4</sup> en fonction des différents facteurs de risque. Une régression de ces valeurs est ensuite effectuée afin d'obtenir une forme paramétrique du passif. Grâce à cette fonction paramétrique du *Best Estimate*, il est alors possible de reconstruire la distribution complète des Fonds Propres en  $t = 1$ , puis de calculer une estimation du SCR par le quantile à 0.5% des Fonds Propres en  $t = 1$ .

Le schéma ci-dessous, extrait d'une documentation interne à Generali, résume le

4. Cette valeur de *Best Estimate* est dite inexacte du fait du nombre très faible de simulations Monte Carlo effectuées lors du calcul. En effet, comme le montre la loi faible des grands nombres, une moyenne converge vers une espérance seulement lorsque le nombre de simulations tend vers l'infini.

principe de l'approche LSMC pour l'estimation du capital économique d'une compagnie d'assurance par la projection à 1 an du passif.

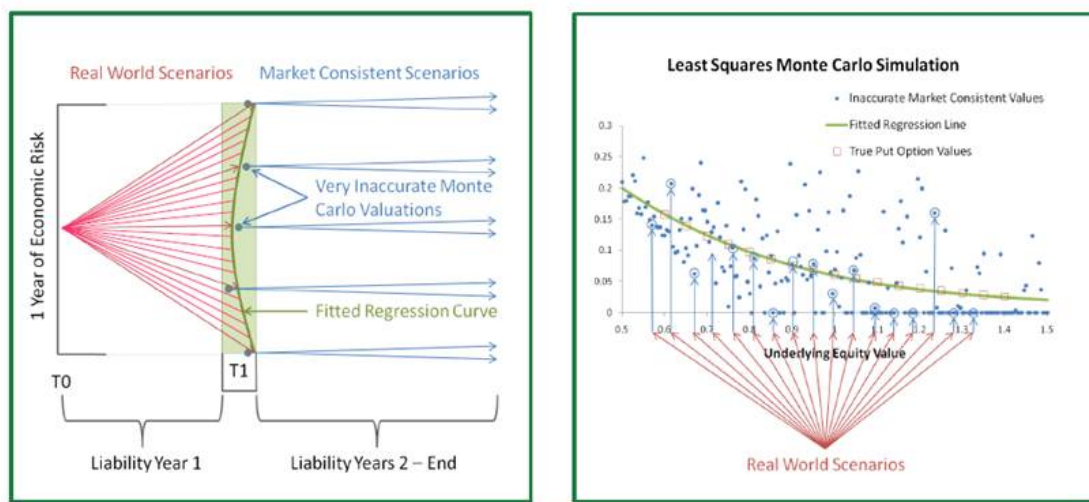


FIGURE 2.4 – Schématisation de l'approche *Least Square Monte Carlo*

Source : Documentation interne à Generali sur la méthodologie du modèle interne

## 2.3 Approche adoptée par Generali

Cette section a pour objectif d'explicitier la manière dont est estimée la distribution des passifs stressés de Generali Vie via leur modèle interne, dans une optique de calcul de SCR.

### 2.3.1 Caractéristiques de la ligne d'activité étudiée

Au cours de cette étude, nous nous sommes concentré sur une ligne d'activité particulière du portefeuille de Generali Vie. Nous appellerons cette ligne d'activité (LoB) la LoB unique d'assurance vie. Elle regroupe tous les contrats d'assurances vie qui présentent des interactions actifs/passifs. Cette ligne d'activité est essentiellement composée d'épargne, mais elle rassemble également des contrats de retraite, de protection.

Tous les résultats présentés dans le corps du texte auront été calculés sur cette ligne d'activité. Certains résultats obtenus sur d'autres lignes d'activité sont présentés en Annexe C.

### 2.3.2 Identification des risques

La première étape de calcul du capital économique requis de Generali, et plus particulièrement de la projection de la distribution des passifs *Best Estimate* par LSMC,

consiste à identifier l'ensemble des facteurs de risque qui pourraient impacter le passif du bilan prudentiel sur une période d'un an.

Dans le cadre du Modèle Interne, Generali a identifié un certain nombre de risques en étudiant l'ensemble des garanties de leurs contrats d'assurances et des risques sous-jacents à ces contrats. L'ensemble de ces risques a été validé par le processus d'approbation du Modèle Interne. L'identification de ces différents risques a été faite par Generali préalablement à cette étude et n'a pas été remise en cause ici. Ceux-ci nous ont été fournis directement afin de servir de base d'apprentissage à nos modèles.

La ligne d'activité étudiée en détail dans ce mémoire, la LoB unique d'assurance vie, regroupe tous les contrats d'assurance vie présentant des interactions ALM. Pour cette ligne d'activité, voici les risques qui ont été identifiés par Generali et approuvés par l'ACPR, en fonction desquels sera exprimée la fonction paramétrique du *Best Estimate* choqué.

Les différents facteurs de risque sont corrélés entre eux à partir d'une matrice de corrélation définie positive<sup>5</sup> et symétrique<sup>6</sup> définie par une copule<sup>7</sup> gaussienne.

### Facteurs de risque de souscription

- **Mortalité** : le risque de mortalité est défini comme le risque de perte, ou de changement défavorable de la valeur des passifs d'assurance, résultant d'un changement de niveau, de tendance ou de volatilité des taux de mortalité, où une augmentation du taux de mortalité amènerait à une augmentation de la valeur du passif assurantiel. Pour traduire l'effet sur les passifs de ces trois composantes de la mortalité, nous introduisons un unique facteur de risque : **Mort**.
- **Longévité** : le risque de longévité est quant à lui le risque dû à un changement de niveau, tendance ou volatilité du taux de mortalité où une diminution du taux de mortalité conduirait à une augmentation de la valeur des passifs. Pour traduire l'effet sur les passifs de ces trois composantes de la longévité, nous introduisons un unique facteur de risque : **Long**.
- **Mortalité CAT** : le risque de mortalité catastrophique est le risque de perte ou de l'augmentation soudaine de la mortalité au cours de la première année de projection en raison d'une pandémie, d'un acte de terrorisme, d'une catastrophe

---

5. Une matrice  $A$  est dite définie positive si et seulement si, pour toute matrice colonne  $X$  non nulle,  $X^t A X > 0$ .

6. Une matrice de taille  $N \times M$ ,  $A = (a_{i,j})_{(i,j) \in N \times M}$  est dite symétrique si et seulement si  $\forall (i,j) \in N \times M$ ,  $a_{i,j} = a_{j,i}$ .

7. Une copule est fonction du cube unité  $C : [0, 1]^d \rightarrow [0, 1]$  qui satisfait les propriétés suivantes :

- Soit  $x \in [0, 1]^d$ ,  $C(x) = 0$  si il existe  $i \in [1, d]$  tel que  $x_i = 0$ .
- Soit  $x \in [0, 1]^d$ ,  $C(x) = 1$  si  $\forall i \in [1, d]$   $x_i = 1$ .
- $C$  augmente de façon monotone  $\forall x_i \in [0, 1]^d$ .

naturelle. Pour traduire l'effet sur les passifs de ce risque, nous introduisons un unique facteur de risque : **MCat**.

- **Rachat** : le risque de rachat est le risque de perte ou de variation défavorable de la valeur des passifs d'assurance résultant de variations du niveau ou de la volatilité des taux de rachat, de résiliations, et de renouvellements des polices d'assurances. Ce risque est inhérent aux options contractuelles permettant aux assurés de résilier totalement ou partiellement, de racheter, de restreindre ou suspendre leur couverture d'assurance. Le facteur de risque de rachat sera noté **Lapse**.
- **Frais** : le facteur de risque de frais est le risque de perte, ou de changement défavorable de la valeur des passifs d'assurance, résultant d'un changement de niveau, de tendance ou de volatilité des frais engagés pour la gestion des contrats d'assurance ou de réassurance (risque de frais de vie). Il sera noté **Expe**.

### Facteurs de risques financiers

Les différents facteurs de risques financiers sont calibrés à partir de données historiques.

- **Taux d'intérêt** : Le risque découlant de la variation des taux d'intérêt est complexe à prendre en compte étant donné que la courbe des taux est composée de plusieurs points. Pour cette raison, ce risque est scindé en 3 facteurs de risque déterminés par une analyse en composantes principales : **IR\_PC1**, **IR\_PC2** et **IR\_PC3**. Ils représentent respectivement le risque lié à un changement de niveau (à l'origine de déformations parallèles de la courbe), de pente et de courbure (à l'origine de changement de convexité) de la courbe de taux.

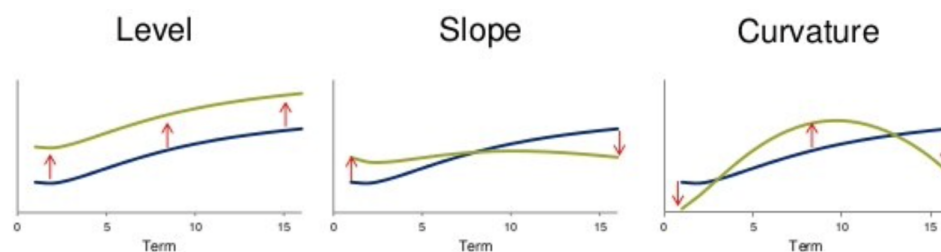


FIGURE 2.5 – Différentes déformations de la courbe des taux prises en compte comme facteurs de risque

Afin de construire l'ensemble de données contenant les chocs sur les composantes principales à partir des séries temporelles des taux d'intérêts, les étapes suivantes sont effectuées :

- Dans un premier temps les données sont examinés et le traitement des points de données manquants est défini par interpolation et extrapolation conformément à la méthodologie Smith-Wilson.
- Dans un second temps, les variations logarithmiques mensuelles historiques ajustées du seuil des taux d'intérêt sont calculées.

$$\Delta r_t = \ln\left(\frac{r_t - s}{r_{t-1} - s}\right)$$

Avec  $r_t$  le taux d'intérêt en date  $t$  et  $s$  le taux d'intérêt le plus bas possible (paramètre seuil).

- Enfin, une analyse en composantes principales (ACP) sur le logarithmique mensuel ajusté au seuil calculé précédemment est effectuée.

Un autre risque pris en compte est celui inhérent à un changement de volatilité des taux d'intérêt : **IRIV**.

Le risque de capital inhérent à la volatilité du taux d'intérêt est défini comme le pire changement des fonds propres de base à 1 an, dû à une variation du niveau de volatilité du taux d'intérêt. Ce facteur de risque est modélisé par une approche de Point d'ancrage avec facteurs d'échelle de volatilité.

Cette approche commence par le calibrage d'une distribution pour un seul point sur la surface de volatilité des taux d'intérêt (le « point d'ancrage »). Les chocs sur d'autres points de la surface sont ensuite dérivés dans la simulation en multipliant le choc du point d'ancrage par des facteurs d'échelle spécifiques à la durée des options et à la maturité des swaps qui reflètent l'évolution de la surface de volatilité des taux d'intérêts au fil du temps.

Pour les volatilités des taux d'intérêt, Generali utilise la surface de volatilité des taux d'intérêt basée sur les volatilités implicites des swaptions à la monnaie (ATM) comme entrée et référence. Et le point d'ancrage choisi est la volatilité des taux d'intérêt sur 5 ans par 5 ans.

- **Action** : Le risque action reflète les conséquences d'une variation du marché actions sur la valeur de marché du portefeuille d'actions de Generali. Deux facteurs de risque sont pris en compte : **Idx\_Eq** reflétant la baisse du prix des actions et **EQIV** quant à lui un facteur lié à un changement de volatilité des marchés des actions.

Le facteur de risque  $Idx\_Eq$  est déterminé par le log-rendement des valeurs du portefeuille actions :

$$Idx\_Eq = \frac{VM_{action\ t=0^+} - VM_{action\ t=0}}{VM_{action\ t=0}}$$

Il est supposé que ce facteur de risque suit une loi de Student.



Quant à lui le facteur de risque EQIV est calibré par l'approche de point d'ancrage avec facteurs d'échelle de volatilité. Cette approche a été détaillée pour la modélisation du facteur de risque EQIV.

- **Obligation** : Le facteur de risque **IdxBond**, reflète les conséquences d'une variation du marché obligations sur la valeur de marché du portefeuille obligataire de Generali.

Il est déterminé de la même manière que  $Idx\_Eq$ , c'est à dire à partir du log-rendement de la valeur de marché du portefeuille obligataire :

$$IdxBond = \frac{VM_{oblig\ t=0+} - VM_{oblig\ t=0}}{VM_{oblig\ t=0}}$$

et est supposé suivre une loi de Student.

- **Immobilier** : Le risque inhérent aux conséquences d'une variation du marché immobilier sur la valeur de marché du portefeuille d'actif de Generali : **Idx\_Pr**.

Il représente le log-rendement des valeurs du portefeuille immobilier :

$$Idx\_Pr = \frac{VM_{immo\ t=0+} - VM_{immo\ t=0}}{VM_{immo\ t=0}}$$

Il est supposé que ce facteur de risque suit une loi de Student.

- **Volatility Adjustment** : le facteur représentant la variation de la *Volatility Adjustment* : **dVA**. dVA correspond à un delta de *Volatility Adjustment*, c'est à dire :

$$dVA = VA_{0+} - VA$$

**NB** : Le risque de variation de l'ajustement pour volatilité n'est pas à proprement parler un risque puisqu'il est inhérent au risque de crédit. En effet, le *Volatility Adjustment* est une mesure de Solvabilité 2 destinée aux branches longues, qui vise à ajuster la volatilité de la solvabilité d'une compagnie d'assurance afin de la rendre moins sensible aux conditions économiques. C'est un facteur d'ajustement, fonction des *spreads* obligataires et des devises d'un portefeuille de référence EIOPA, qui va être ajouté à la courbe des taux lors de l'actualisation des cash-flows de passifs et des actifs obligataires. Il va réduire la sensibilité des Fonds Propres à la volatilité des taux courts.

La *Stochastic Volatility Adjustment* est introduit dans le modèle interne afin de prendre en compte le changement de l'ajustement de volatilité si le choc de crédit change, et ainsi d'avoir un ajustement de volatilité différent pour chaque simulation.

### Construction des scénarios stressés en $t=0^+$ à partir des facteurs de risque

Afin de construire les scénarios de chocs utilisés dans l'approche *Least Square Monte Carlo* pour accéder à la distribution complète du *Best Estimate* stressé, et in fine à la valeur du SCR, des chocs sur les hypothèses économiques centrales vont être effectués à partir des facteurs de risques identifiés.

Tout d'abord, un bilan central est calculé en  $t=0$  à partir du modèle de projection ALM ainsi que d'hypothèses sous-jacentes centrales. En  $t = 0^+$ , ce bilan économique central va être choquer par les différents facteurs de risques de la manière suivante :

- Les valeurs de marché des portefeuilles action, obligation et immobilier seront obtenus par la formule suivante :

$$VM_{t=0^+} = VM_{t=0} \times (1 + e^{Idx})$$

avec VM la valeur de marché et Idx respectivement Idx\_Eq, IdxBond et Idx\_Pr.

- Le spread de *Volatility Adjustment* sera augmenté ou diminué du choc dVA.

$$VA_{shocked} = VA_{Base} + dVA$$

- Les taux d'intérêt en scénario de stress sont obtenus par la formule suivante :

$$r_{shocked} = (r_{Base} - s) \times e^{shock} + s$$

avec

- $r_{shocked}$  la courbe de taux choquée ;
- $r_{Base}$  la courbe de taux centrale ;
- $shock$  le choc appliquée à la courbe de taux. Ce choc est obtenu par une combinaison linéaire des vecteurs propres multipliés par les composantes principales (CP) choquées. C'est à dire  $Shock = EVec^1 \times PC1 + EVec^2 \times PC2 + EVec^3 \times PC3$ , avec EVec les vecteurs propres associés.

- La volatilité des taux d'intérêt est choquée de la manière suivante :

$$IV_{Shocked}^{O|S} = IV_{BaseCase}^{O|S} + Z$$

Avec

- $IV_{Shocked}^{O|S}$  la volatilité des taux d'intérêt choquée pour les points de la surface de volatilité des taux d'intérêt avec une durée d'option O et une durée de swap S.
- $IV_{BaseCase}^{O|S}$  la volatilité des taux d'intérêt centraux pour les points de la surface de volatilité des taux d'intérêt avec une durée d'option O et une durée de swap S.

- Z est la variation du niveau de volatilité des taux d'intérêt dans la simulation Monte-Carlo (le facteur de risque).
- De même, la volatilité des actions est choquée par l'approche de point d'ancrage, et plus précisément par :

$$EV_{Shocked}^T = EV_{BaseCase}^T \times e^Z$$

Avec

- $EV_{Shocked}^T$  la volatilité implicite des actions choquée pour une durée d'option T.
- $EV_{BaseCase}^T$  la volatilité implicite des actions centrale pour une durée d'option T à la date de clôture.
- Z est la variation du niveau de volatilité dans la simulation Monte-Carlo (le facteur de risque).
- Le taux de mortalité central  $q_x$  sera augmenté ou diminué par le choc de mortalité, de longévité et de mortalité catastrophique.
- Le choc lié au risque de rachat est appliqué par un stress proportionnel sur les hypothèses centrales.
- Le risque de dépenses est pris en compte en appliquant un vecteur d'inflation stressé reflétant un niveau de sécurité spécifique (par exemple le quantile à 99,5%). Ainsi, le stress inflationniste est calibré séparément pour chaque année civile.

$$Shocked\ Expenses_i = \frac{BE\ Expenses}{\prod_{j=1}^i (1 + \pi_{BE_j})} \times \sum_{j=1}^i (1 + \pi_{stress_j})$$

Avec

- i l'année calendaire,
- $BE\ Expenses$  la meilleure estimation des flux de dépenses futurs,
- $\pi$  le taux d'inflation à terme sur 1 an en l'année j.

### 2.3.3 Projection de la distribution du *Best Estimate* stressé par LSMC 2

Generali calcul son SCR par modèle interne et pour cela utilise l'approche de projection de la distribution du *Best Estimate* stressé, pour ensuite en déduire la distribution des Fonds Propres stressés, et ainsi calculer le SCR.

## LSMC 2 :

Pour cela, Generali adopte ce qu'ils appellent la méthode LSMC 2, qui est un dérivé de la méthode *Least Square Monte Carlo* (LSMC) présentée dans la section précédente. Le nombre de scénarios primaires utilisé est 25 000. Ce qui différencie le LSMC 2 de la méthode LSMC classique est la manière dont sont projetés les facteurs de risque non-financiers. En effet, les facteurs de risques financiers seront simulés 25 000 fois chacun alors que les facteurs de risques non-financiers (comme le risque de rachat, de longévité etc...) sont simulés 200 fois puis chacune des simulations est dupliquée 125 fois.

Ceci est un choix opérationnel résultant du fait que Generali utilise l'approche dite de *Flexing*<sup>8</sup>. Il y donc un modèle déterministe de projection des flux de passifs (Prophet Passif) à part du modèle ALM, déterminant les flux de passifs sans prise en compte de l'impact des interactions actifs/passifs, qui sont par la suite intégrées en *input* du modèle ALM (Prophet ALS). L'utilisation de l'approche LSMC 2 permet donc une économie de temps de calcul en évitant de faire tourner 25 000 fois le modèle déterministe.

### Application des chocs au temps $t=0^+$

Nous rappelons que de manière opérationnelle dans le modèle interne de Generali, les scénarios primaires correspondent à des chocs sur les facteurs de risque appliqués en  $t = 0^+$  et non pas à 1 an.

### Calibration des scénarios primaires et des simulations secondaires

Après avoir identifié et répertorié les risques auxquels est exposé le passif du portefeuille, il faut obtenir les valeurs approchées des *Best Estimate* afin de pouvoir calibrer la forme paramétrique par une régression de ces valeurs sur les facteurs de risque. Pour cela il est nécessaire de procéder à deux niveaux de simulations, comme il a été expliqué dans la section présentant l'approche LSMC.

25 000 scénarios de stress sur les facteurs de risque sont utilisés pour le *fitting* de la régression par l'approche LSMC. Pour sélectionner ces scénarios de manière à ce qu'ils soient le mieux répartis dans l'espace, les caractéristiques des suites SOBOLS à faible divergence sont utilisées afin qu'ils soient uniformément distribués et bien répartis dans l'espace des facteurs de risques.

Pour chacun des scénarios primaires, 2 simulations secondaires sous la probabilité risque neutre sont ensuite réalisées. Afin d'augmenter la convergence de l'approche, les deux simulations secondaires effectuées sont antithétiques.

Le budget total de simulation permettant de calibrer la fonction proxy des provisions *Best Estimate* à 1 an en fonction des facteurs de risque est donc égale à 50 000 simulations.

---

8. Le flexing consiste à projeter les engagements de l'assureur (passifs) dans un modèle déterministe différents du modèle de projection des actifs. Les principaux résultats de ce modèle déterministe sont ensuite intégrés dans le modèle ALM afin de prendre en compte les interactions actif/passif.

### Choix des fonctions de base de la régression

Après avoir réalisé 25 000 scénarios primaires correspondant à 25 000 chocs sur les différents facteurs de risque, puis pour chacun d'entre eux 2 simulations antithétiques sous la probabilité risque neutre, des estimateurs peu fiables (dû au faible nombre de simulations Monte Carlo effectuées) du *Best Estimate* sont alors calculées pour chaque scénario primaire par la moyenne des sommes des flux futurs actualisés obtenues sur les deux simulations secondaires. Les approximations des provisions *Best Estimate* stressées vont alors servir de base pour effectuer une régression de ces valeurs en fonction des facteurs de risque, dans le but d'obtenir une fonction paramétrique définissant le *Best Estimate* stressé en fonction des facteurs de risques.

#### *Régression polynomiale*

Le choix des fonctions de base utilisées dans la régression doit prendre en compte deux phénomènes :

- **Non-linéarité** : le montant de l'engagement de l'assureur n'est généralement pas linéaire avec le stress des facteurs de risque. Le doublement d'un stress en un facteur de risque ne se traduit nécessairement pas par un doublement exact des engagements. Pour prendre en compte ce phénomène, Generali intègre à la régression les puissances des facteurs de risque.
- **Non-additivité** : l'impact de deux risques survenant en même temps est souvent très différent de la somme des impacts calculés individuellement. Afin de capter les interactions entre les facteurs de risque, le modèle de régression de Generali intègre les termes croisés des puissances des facteurs de risque.

Finalement, Generali a donc choisi comme fonction de base de la régression des termes polynomiaux de chacun des facteurs de risque identifiés ainsi que leurs termes croisés. Le degré maximal des monômes déterminé par Generali est 4 et diffère en fonction de la LoB (*Ligne of Business*) étudiée, et le degré maximal d'un terme croisé<sup>9</sup> est de 4.

La limite de la régression polynomiale est le nombre de termes qu'elle engendre. En effet, dans le cas de 15 facteurs de risque et d'un degré maximal autorisé de 3, le modèle contiendrait  $15 \times 3 = 45$  monomes. En ajoutant à cela les différents termes croisés, cela amène à un nombre excessif de termes. Afin de réduire le nombre de termes présents dans la régression polynomiale tout en optimisant la qualité du modèle, un algorithme de sélection des termes *Adaptive Stepwise Forward* est réalisé.

#### *Algorithme de sélection des termes : Adaptive Stepwise Forward*

Les termes polynomiaux des 15 facteurs de risque identifiés ainsi que les termes croisés introduisent un nombre de termes considérable dans la régression. Afin de maximiser la

---

9. Soit  $x_k^i x_l^j$  un terme croisé, le degré de ce terme est alors égale à  $i+j$ .

performance de la régression établie tout en minimisant sa complexité, un algorithme de sélection des termes est appliqué : l'*Adaptive Stepwise Forward* (Cf ci-dessous 2.3.3).

Cet algorithme fonctionne de la manière suivante :

1. Un modèle contenant seulement l'intercept est élaboré, puis son AIC (*Akaike Information Criterion*, critère de performance) est calculé.
2. Chacun leur tour, les termes de degré 1 sont ajoutés au modèle contenant seulement l'intercept, puis leur AIC est calculé. Le terme ayant diminué le plus l'AIC est alors ajouté au modèle.
3. À n'importe quelle étape du processus de sélection, un terme n'est considéré pour inclusion dans le modèle que si le modèle contient déjà tous les facteurs du nouveau terme. Par exemple, considérons trois facteurs de risque  $z_1$ ,  $z_2$  et  $z_3$ . A un stade donné de la sélection, en supposant que le modèle est donné par :

$$y = \beta_{0,0,0} + \beta_{1,0,0}z_1 + \beta_{0,1,0}z_2 + \beta_{2,0,0}z_1^2$$

Les termes marginaux pouvant être ajoutés au modèle seraient alors :

$$\{z_3, z_1z_2, z_1^3, z_1^2z_2 \text{ et } z_2^2\}$$

Les termes candidats sont ajoutés séparément au modèle un à un, le terme diminuant le plus l'AIC est ajouté au modèle.

4. Ce processus est répété en ajustant l'ensemble des termes candidats à chaque étape, jusqu'à une condition d'arrêt. Les différentes conditions d'arrêt sont les suivantes : le nombre maximal de termes dans le modèle<sup>10</sup> a été atteint, aucun terme candidat ne diminue l'AIC du modèle où tous les termes d'un degré inférieur au degré maximum établi ont été considérés.

#### *AIC : Akaike Information Criterion*

L'AIC ou Critère d'Information d'Akaike, utilisé pour le choix des termes à garder dans la régression, est basé sur la vraisemblance du modèle et permet de pénaliser l'ajout d'un terme dans le modèle. La vraisemblance est définie en fonction d'un vecteur de paramètres  $\theta$  comme la densité des données observées par rapport à une mesure de probabilité discrète ou continue.

$$L(\theta|x_1, \dots, x_n) = \prod_{i=1}^n L(\theta|x_i) = \prod_{i=1}^n f_{\theta}(x_i)$$

avec  $f_{\theta}$  la densité de probabilité dépendant de  $\theta$ .

Ce critère s'écrit comme suit :

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

avec

---

10. nombre déterminé par l'utilisateur

- $k$  le nombre de paramètres du modèle à estimer
- $L$  le maximum de vraisemblance du modèle

Il s'agit de trouver un compromis entre la qualité de l'ajustement et la complexité du modèle, en évitant ainsi un sur-ajustement.

Le critère d'information d'Akaike est très utilisé pour comparer des modèles de différentes tailles les uns avec les autres, le meilleur modèle sera alors celui dont l'AIC sera le plus faible.

### 2.3.4 Obtention de la distribution complète des Fonds Propres stressés

À partir de la fonction proxy obtenue suite à la régression du LSMC, la distribution complète des engagements *Best Estimate* stressés en  $t=0^+$  est obtenue en appliquant cette fonction à  $2^{18}$  (soit 262 144) scénarios de facteurs de risque stressés. Ces  $2^{18}$  scénarios sont répartis uniformément dans l'espace des facteurs de risque par des suites Sobols, et constituent une table appelée la *Risk Factor Scenario Set* (RFSS). L'approche *Least Square Monte Carlo* nous permet donc d'obtenir l'approximation suivante de la distribution complète du *Best Estimate* stressé en  $0^+$  :

$$BE_{stress\ t=0^+} \approx f(RF)$$

avec :

- $f$  la fonction paramétrique obtenue par la régression de l'approche *Least Square Monte Carlo*.
- RF le vecteur des facteurs de risque considérés.

La distribution complète des Fonds Propres stressés est alors déduite de la distribution complète du *Best Estimate* stressé en soustrayant aux actifs les passifs :

$$FP_{t=0^+} = Actifs_{t=0^+} - (BE_{t=0^+} + OBSI_{t=0^+}) \quad (2.1)$$

avec :

- $FP_{t=0^+}$  les Fonds Propres stressés en  $t=0^+$
- $Actifs_{t=0^+}$  les actifs en valeurs de marché stressés en  $t=0^+$ , valorisés par la méthode Curve Fitting.
- $BE_{t=0^+}$  le *Best Estimate* stressé à  $t = 0^+$ , calculé comme décrit précédemment par la méthode LSMC 2.
- $OBSI_{t=0^+}$  les autres postes du bilan (*Other Balance Sheet Item*), non mentionnés dans le bilan simplifié.

### 2.3.5 Déduction du capital de solvabilité requis (SCR)

Enfin, le capital économique de solvabilité requis peut se calculer à partir des Fonds Propres initiaux et de la distribution complète des Fonds Propres à 1 an par la formule suivante :

$$SCR_{t=0} \approx FP_{t=0} - q_{0.5\%}(FP_{stress \text{ en } t=0+}) \quad (2.2)$$

Les Fonds Propres initiaux sont obtenus en soustrayant les postes du passif du bilan prudentiel aux postes de l'actif à l'instant  $t=0$ . Pour rappel, afin de valoriser le poste de provisions *Best Estimate* au temps  $t=0$ , il est nécessaire de procéder par simulations de Monte Carlo. C'est-à-dire que, connaissant l'environnement économique initial, les *cash flows* de passifs vont être projetés à peu près 1000 fois sur un horizon de temps défini (environ 30 ans) en univers risque neutre grâce à un modèle ALM permettant de capter les interactions entre l'actif et le passif. Puis la valeur du *Best Estimate* initial s'obtient par l'estimateur suivant :

$$\hat{BE} = \frac{1}{N} \sum_{i=1}^N [CF_i | Y_i]$$

avec

- CF les *cash flows* actualisés.
- $Y_i$  l'environnement économique au temps  $i$ .

La distribution complète des Fonds Propres stressés en  $t=0^+$  s'obtient quant à elle par la formule (2.1) ou plus précisément par l'approximation de LSMC suivante :

$$FP_{t=0+} \approx Actifs_{t=0+} - (f(RFSS) + OBSI_{t=0+}) \quad (2.3)$$

avec

- $f$  la fonction paramétrique du *Best Estimate* stressé.
- RFSS la *Risk Factor Scenario Set*.



## CONCLUSION DU CHAPITRE :

Dans le cadre d'une approche Monte Carlo complète, le calcul du capital économique pour une compagnie d'assurance vie ayant un modèle ALM complexe nécessite des calculs stochastiques imbriqués afin de capter la valeur temps des options et garanties présentes dans les contrats. Cela implique l'exécution de plusieurs milliers de simulations primaires (monde réel) projetant les conditions économiques à 1 an, suivi d'un grand nombre de simulations secondaires (risque neutre) afin de calculer une valeur fiable de la provision *Best Estimate* liée à chacun des scénarios primaires. Cette approche stochastique imbriquée, nommée Simulations dans les Simulations, nécessite un temps de calcul déraisonnable voire infaisable opérationnellement tant la charge de calcul est importante.

Différentes approches ont vu le jour afin de résoudre le problème de temps de calcul inhérent aux Simulations dans les Simulations. Nous pouvons citer l'accélérateur de SdS ; l'approche de *Replicating Portfolio*, l'approche *Curve Fitting* ainsi que l'approche *Least Square Monte Carlo*.

Parmi ces différentes approches, Generali a choisi de valoriser ses passifs d'assurance vie par une approche LSMC (dans laquelle la distribution des passifs est calculée en  $t = 0^+$  et non pas  $t=1$  an), appliquée avec 25 000 simulations primaires suivies de 2 simulations secondaires (antithétiques). La régression réalisée dans cette approche en vue d'obtenir une fonction paramétrique du BE en fonction des facteurs de risques est une régression polynomiale, contenant des termes croisés de plusieurs facteurs de risque. Les termes de cette régression polynomiale sont rigoureusement sélectionnés par un algorithme *Adaptive Forward Stepwise* afin d'en maximiser les performances.

## TRANSITION :

*Serait-il possible d'améliorer cette étape de régression de la méthode LSMC par un algorithme de Machine Learning ?*

*À travers le chapitre suivant, différents modèles de régression de machine learning seront introduits, puis appliqués à la régression des Best Estimate obtenus par l'approche LSMC, dans le but de les confronter à la régression polynomiale appliquée actuellement dans le modèle interne de Generali.*



## Chapitre 3

# Utilisation de modèles de *Machine Learning* pour défier l'étape de régression du LSMC

### *Qu'est-ce que le Machine Learning ?*

"*Machine Learning is the science of getting computers to learn without being explicitly programmed*" Arthur Samuel, pionnier de l'apprentissage automatique, 1959.

En effet, l'apprentissage automatique ou *Machine Learning* en anglais, consiste à laisser l'ordinateur apprendre quel calcul effectuer, plutôt que de lui donner ce calcul (c'est-à-dire le programmer de façon explicite). Pour ce faire, les méthodes d'apprentissage automatique sont fortement inspirées de la façon dont les humains apprennent. Parmi ces méthodes, les deux plus connues sont : l'apprentissage supervisé et l'apprentissage non-supervisé. Un algorithme d'apprentissage supervisé va chercher à trouver le lien existant entre une variable cible  $Y$  et  $p$  variables explicatives  $X = X_1, \dots, X_p$ , en apprenant à partir d'une base de données contenant une multitude d'exemples de réponses  $Y_n$  en fonction de réalisations de  $X^n = X_1^n, \dots, X_p^n$ . Quant à lui, l'algorithme d'apprentissage non-supervisé ne connaît pas d'exemple de réponse, mais va analyser une base de données afin de trouver des règles générales liant les données entre elles.

Nous nous intéresserons seulement à l'apprentissage supervisé au cours de ce mémoire, étant donné que c'est l'approche utilisée dans les modèles de régressions multivariées qui vont suivre.

Afin d'apprendre une relation entre les variables, l'algorithme d'apprentissage supervisé a besoin en entrée d'un tableau de données d'expériences, contenant pour plusieurs valeurs de la variable cible ou variable d'intérêt  $Y$ , les valeurs des variables explicatives, ou prédicteurs  $X = (X_1, \dots, X_p)$  qui lui sont associées. Le modèle va alors dégager de ce jeu de données d'expérience la relation liant les variables explicatives à la variable cible, pour ensuite être capable de déterminer la valeur de la variable cible lorsqu'on lui donnera seulement de nouvelles valeurs des variables explicatives.

$$Y_n = f(X_1^n, \dots, X_p^n) + \epsilon_n$$

À travers ce chapitre seront présentées dans un premier temps les données utilisées pour l'apprentissage ainsi que la validation des différents modèles, puis dans un second temps sera abordée la théorie derrière les modèles utilisés [6, 18, 24]. Enfin, les résultats obtenus par les modèles seront présentés dans le but de les comparer et de choisir le modèle qui pourrait être le plus adapté à la problématique de ce mémoire.

### *Quel est l'objectif des modèles dans le cas de cette étude ?*

L'utilisation de modèles de *Machine Learning* pour prédire la distribution du *Best Estimate* stressé d'une compagnie d'assurance a déjà été explorée dans certains mémoires d'actuariat. Notamment R. GAUVILLE, dans son mémoire intitulé "*Projection du ratio de solvabilité : des méthodes de machine learning pour contourner les contraintes opérationnelles de la méthode des SdS*" [8], a cherché à expliquer le BE en fonction de différents facteurs de risque par des modèles de *Machine Learning* et obtient des résultats encourageants. La limite de la méthode utilisée dans cette étude est la création de la base de données utilisée pour entraîner les modèles. Ses modèles sont basés sur les résultats obtenus par la méthode de Simulations dans les Simulations, ce qui implique que la base de données utilisée a nécessité des millions de simulations. Lors de la prise en compte d'interactions complexes entre l'actif et le passif, un tel budget de simulations est infaisable d'un point de vue opérationnel.

Pour cette raison, lors de cette étude, les modèles d'apprentissage supervisé ne vont pas s'entraîner sur une base de données construite par un jeu complet de simulations Monte Carlo de type SdS, mais par une approche *Least Square Monte Carlo* effectuant 25 000 simulations primaires puis 2 simulations secondaires antithétiques pour chacune des simulations primaires. Les performances des modèles seront quant à elles testées sur une base de validation de 125 points, construite par un jeu complet de simulations : 125 simulations primaires puis 1 000 simulations secondaires.

## **3.1 Données**

Lors de cette section seront présentées la base de données qui sera entrée en input des modèles d'apprentissage supervisé, appelée *base d'apprentissage*, ainsi que la base de données ensuite utilisée pour la validation des modèles de régression, appelée *base de validation*.

### **3.1.1 Base d'apprentissage**

La base d'apprentissage utilisée contient 25 000 lignes, correspondant aux différents scénarios primaires, et 15 colonnes. Les 14 premières colonnes représentent les prédicteurs du *Best Estimate* de la LoB unique, qui correspondent aux facteurs de risque du modèle

interne et la dernière colonne représente le *Best Estimate* stressé inexacte de la LoB unique<sup>1</sup> obtenue par seulement deux simulations risque neutre.

Les différents facteurs de risque utilisés comme prédicteurs du *Best Estimate* de la LoB unique d'assurance vie sont les suivants :

- Idx\_Bond : Risque lié à une variation du marché obligations.
- Idx\_Eq : Risque lié à une variation du marché actions.
- Idx\_Pr : Risque lié à une variation du marché immobiliers.
- Lapse : Risque de rachat.
- MCat : Risque de mortalité dû à une pandémie.
- Mort : Risque lié à la mortalité.
- Long : Risque lié à la longévité.
- Expe : Risque lié aux frais.
- EQIV : Risque lié à la volatilité du marché actions.
- IRIV\_EUR : Risque lié à la volatilité du taux d'intérêt.
- IR\_PC1\_EUR : Risque lié au changement de niveau de la courbe des taux d'intérêt.
- IR\_PC2\_EUR : Risque lié au changement de pente de la courbe des taux d'intérêt.
- IR\_PC3\_EUR : Risque lié au changement de convexité de la courbe des taux d'intérêt.
- dVA : Delta de l'ajustement de volatilité.

Ces facteurs de risque ont été calibrés par Generali au préalable, et cela n'a pas fait parti de cette étude. La calibration de ces facteurs de risque a été décrite brièvement dans la section 2.3.2.

La ligne d'activité (LoB) étudiée au cours de cette étude est la LoB unique sur les produits d'assurance vie. Elle rassemble uniquement les contrats d'épargne, de retraite, de protection et de prévoyance présentant des interactions ALM.

Ainsi, la base d'apprentissage permettant d'expliquer le *Best Estimate* de la LoB unique d'assurance vie en fonction des facteurs de risque sera composée de 25 000 points pour 14 variables explicatives.

Pour rappel, l'approche LSMC pour la projection de la distribution du BE stressé consiste à effectuer 25 000 scénarios de stress économique à  $t=0^+$ , puis pour chacun

---

1. Ligne d'activité regroupant les contrats présentant des interactions actif/passif

de ces scénarios primaires deux simulations risque neutre via le modèle de projection ALM pour obtenir une valeur du BE pour chaque scénario. Le BE du scénario  $i$  est alors calculé comme étant la moyenne des deux valeurs de *Best Estimate* obtenues à la suite des simulations secondaires. La construction de cette base d'apprentissage a donc nécessité un budget de 25 000 scénarios primaires x 2 simulations secondaires = 50 000 simulations.

Le choix de cette base d'apprentissage sera discuté dans la conclusion de ce mémoire.

### 3.1.2 Base de validation

La base de validation contient 125 lignes et 15 colonnes. Les colonnes représentent les 14 facteurs de risque ainsi que la valeur de *Best Estimate* obtenue par un jeu complet de simulations (1000), associée à ces facteurs de risque. Contrairement à la base d'apprentissage, les valeurs de BE présentes dans la base de validation sont obtenues par un jeu de simulations complet de Monte Carlo. C'est-à-dire que pour chacun des 125 scénarios primaires, 1 000 simulations secondaires en risque neutre sont effectuées pour obtenir une valeur du *Best Estimate* fiable.

Bien que de dimension inférieure à la base d'apprentissage, cette base de validation a nécessité un budget de simulations bien plus important : 125 scénarios primaires x 1 000 simulations secondaires = 125 000 simulations.

Les 125 scénarios utilisés ont été imposés par l'ACPR afin de valider la performance de la fonction paramétrique du *Best Estimate* obtenue par *Least Square Monte Carlo*.

### 3.1.3 Analyse univariée

Avant de commencer la modélisation il est pertinent d'avoir une idée de l'influence de chacune des variables explicatives sur le *Best Estimate* approché qui est la variable cible. Pour cela, une analyse univariée a été effectuée et est résumée par les figures 3.1 et 3.2 ci dessous. Sur ces graphiques sont représentées par la variable BEGRO les valeurs que prend le *Best Estimate* stressé approché en fonction des valeurs des facteurs de risque, les nuances de bleus représentent la densité des points : plus la couleur est claire et plus il y a de points à cet endroit. La courbe rouge représente quant à elle un lissage des données, permettant de se rendre compte de la tendance générale.

Au regard de ces graphiques, il est d'ores et déjà possible de relever certaines informations quant à l'impact des différents facteurs de risque sur le montant de la provision *Best Estimate*. En effet, il est à noter que  $Idx_{Bond}$ ,  $Idx_{Eq}$ ,  $IR_{PC1\_EUR}$ ,  $dVA$  et  $Idx_{Pr}$  ont l'air d'être les variables les plus impactantes, au regard de la sensibilité du montant de provision *Best Estimate* lors de la variation de ces facteurs de risque.

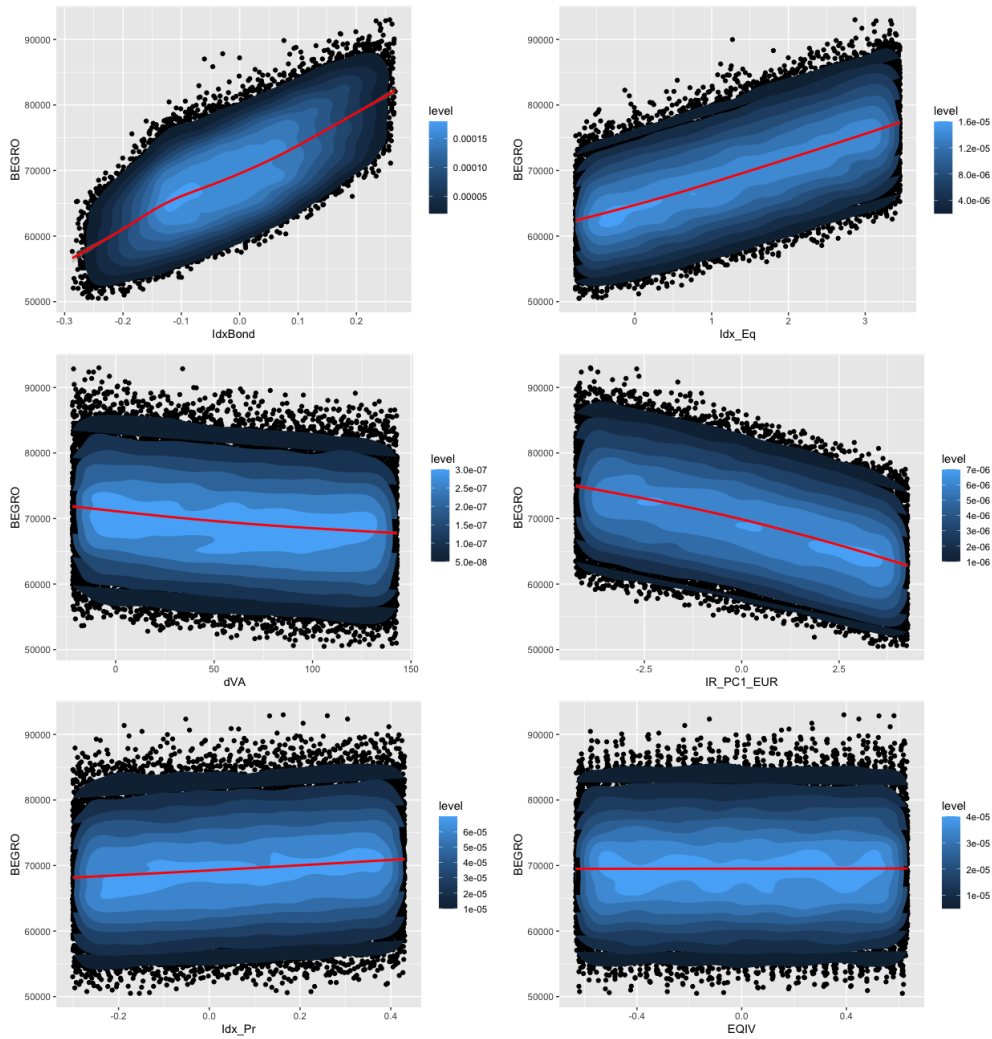


FIGURE 3.1 – Impact de chacun des facteurs de risque majeurs sur le BE individuellement

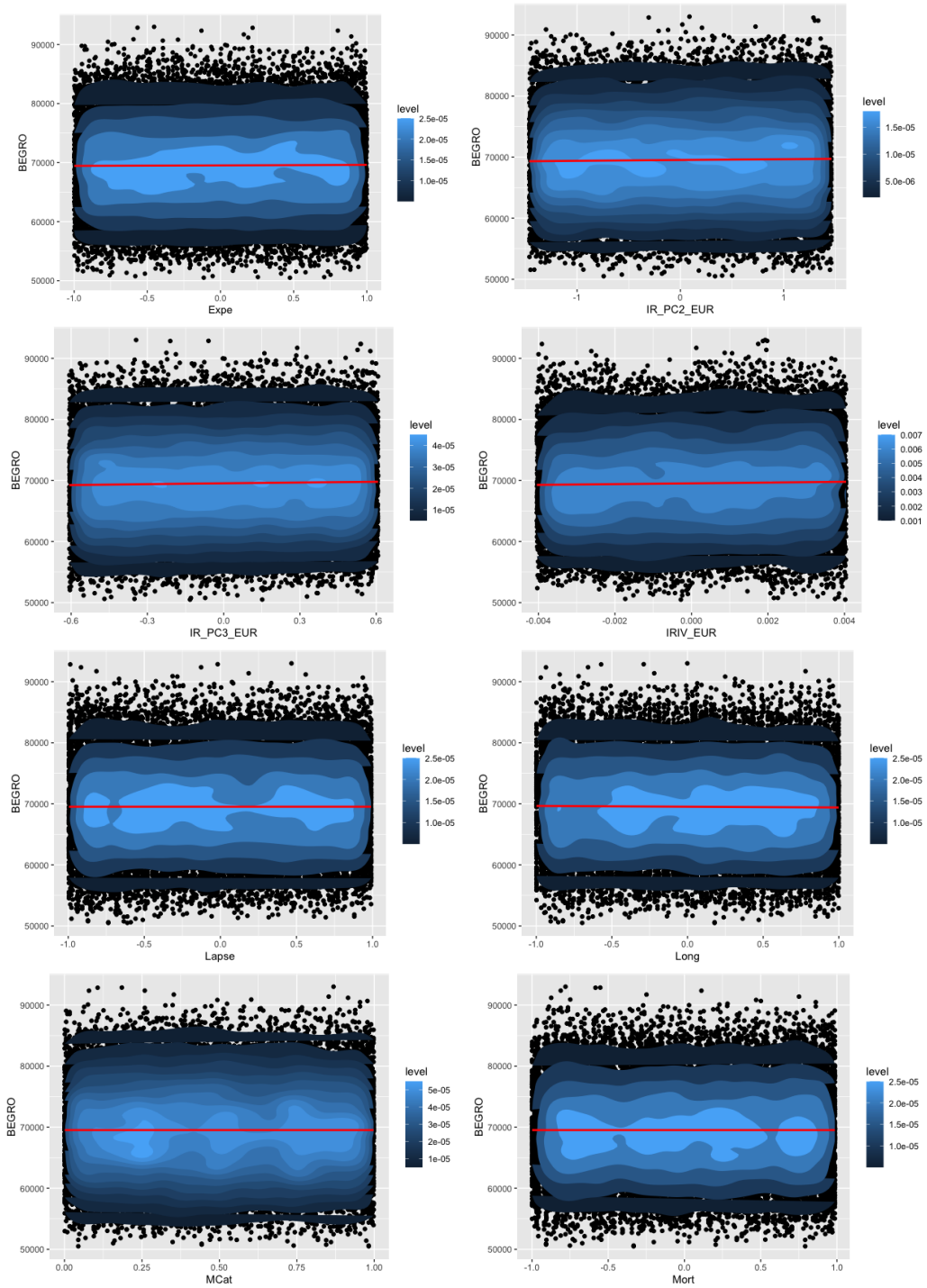


FIGURE 3.2 – Impact de chacun des facteurs risques restants sur le BE individuellement



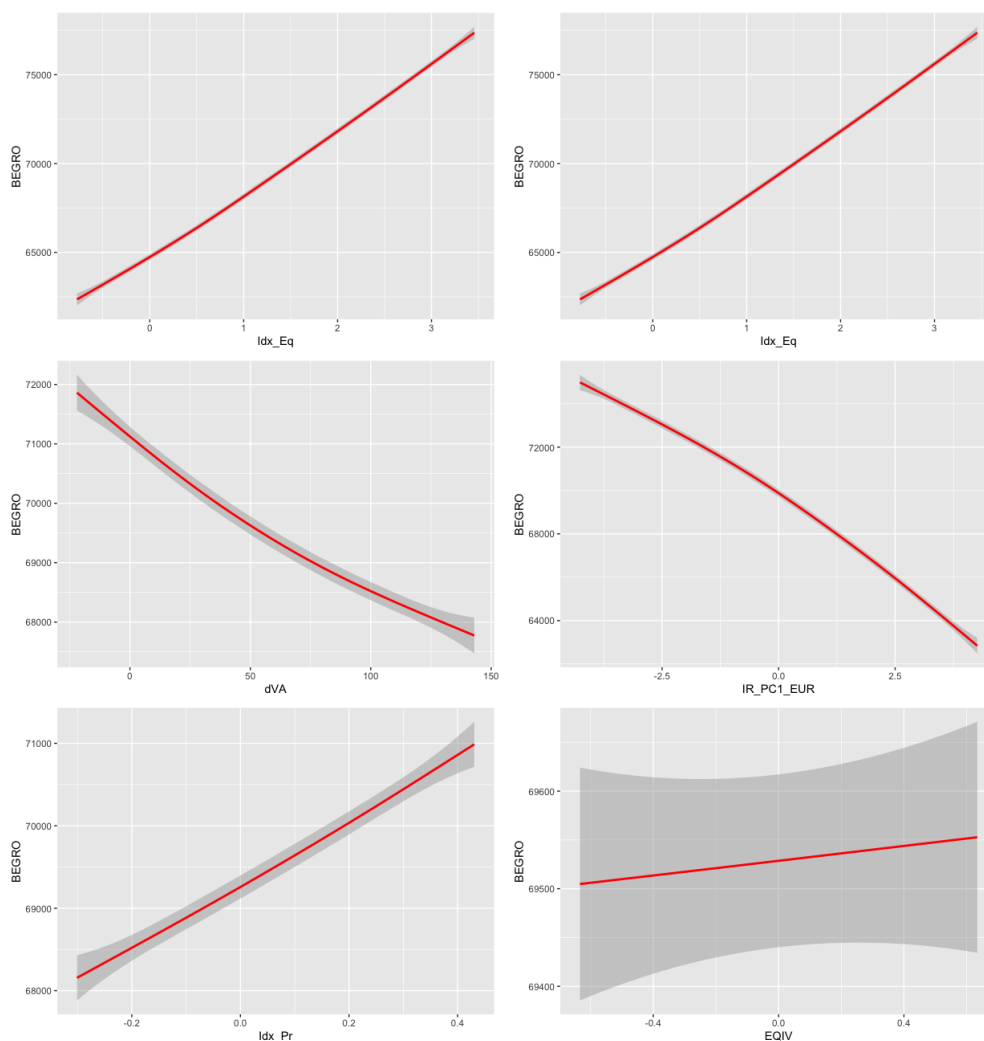


FIGURE 3.3 – Relation liant les facteurs de risque au BE individuellement

Lorsque nous nous intéressons seulement à la fonction de lissage rouge représentée sur les figures 3.3 et 3.4 ci-dessous, liant chacun des facteurs de risque au *Best Estimate*, il ressort que les variables EQIV, Expe, IR\_PC2\_EUR, IR\_PC3\_EUR, IRIV\_EUR ont également un impact notable sur la valeur du *Best Estimate*. Quant à elles, les variables Mort, Lapse et MCat ont l'air d'être peu, voire pas impactantes sur la valeur de la provision *Best Estimate* de la LoB unique.

Il est important de garder cela en tête lors de la construction des différents modèles de régression afin de juger de leur pertinence.

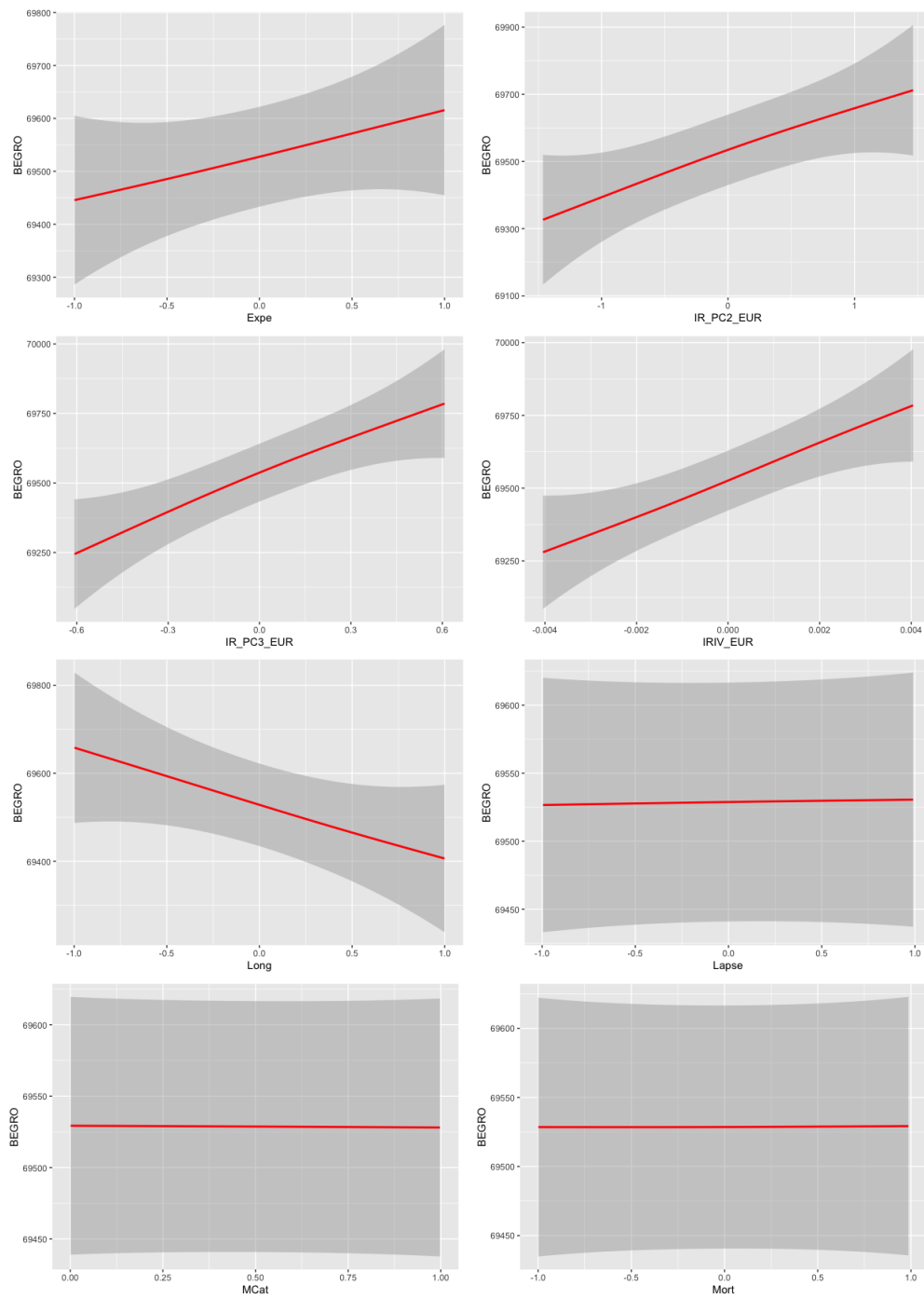


FIGURE 3.4 – Relation liant les facteurs de risque au BE individuellement bis

## 3.2 Théorie des modèles utilisés

Cette section vise à expliquer le fonctionnement des modèles de *Machine Learning* utilisés dans le cadre de ce mémoire pour challenger la régression polynomiale actuellement en place lors de l'approche LSMC.

### 3.2.1 Forêt aléatoire

#### Principe général

Une forêt aléatoire correspond à la réalisation de plusieurs arbres de décision indépendants<sup>2</sup>, en choisissant aléatoirement pour la construction de chacun d'entre eux un certain nombre de variables explicatives. Ces arbres se distinguent également les uns des autres par le sous-échantillon de données sur lequel ils sont entraînés (échantillons bootstrap). Ces sous-échantillons sont tirés au hasard dans le jeu de données initial tout comme les variables explicatives qui serviront à construire l'arbre. Cela a pour but de décorrélérer les arbres de la forêt.

La forêt aléatoire permet de limiter le sur-apprentissage. En effet, l'arbre de décision a tendance à être trop spécifique aux données sur lesquelles il a appris et à mal se généraliser à d'autres bases de données. Or l'objectif d'un modèle de régression est précisément d'être généralisable afin de pouvoir être appliqué à de nouveaux jeux de données pour lesquels on ne connaît pas la valeur de la variable cible : ici le *Best Estimate* stressé. Il faut donc qu'il puisse être généralisable et pas seulement qu'il prédise extrêmement bien les valeurs du *Best Estimate* des individus à partir desquels il a appris (base d'apprentissage). Le fait que la forêt soit construite avec des arbres indépendants la rend mieux généralisable.

Voici les principales étapes de la construction d'une forêt aléatoire :

1. On prend un nombre  $X$  d'observations du jeu de données de départ (avec remise).
2. On prend un nombre  $K$  des  $M$  variables disponibles.<sup>3</sup>
3. On entraîne un arbre de décision sur ce jeu de données.
4. On répète les étapes 1. à 4.  $N$  fois de sorte à obtenir  $N$  arbres.

Pour une nouvelle observation pour laquelle nous cherchons la valeur de la variable cible  $Y$ , il suffit de descendre les  $N$  arbres. Chaque arbre de décision construit prédit une valeur pour la variable cible. La valeur retenue pour la prédiction de la forêt est la moyenne des valeurs obtenues parmi tous les arbres de la forêt.

---

2. Le fonctionnement d'un arbre décisionnel est détaillé en Annexe A

3. Par exemple, si  $K=3$  : seulement l'indice action, la mortalité et le rachat pourraient être utilisés pour construire le premier arbre. Puis 3 autres facteurs de risque choisis aléatoirement pour le deuxième, etc.

### Échantillon *Out Of Bag* (OOB)

Comme mentionné ci-dessus, les arbres de la forêt sont construits à partir d'échantillons *bootstrap* différents issus de la base de données d'apprentissage. L'échantillon à partir duquel un arbre est entraîné s'appelle l'échantillon *in-bag*, alors que les individus non-utilisés lors de la construction de l'arbre constituent l'échantillon *out of bag*, noté OOB.

### Importance des variables

Le modèle de forêt aléatoire nous permet d'accéder à l'importance des différentes variables explicatives au sein du modèle, c'est-à-dire le pouvoir de prédiction des variables explicatives. L'importance des variables peut être calculée de deux manières différentes : soit par le *Mean Decrease Impurity* (MDI), correspondant à la réduction d'impureté des noeuds due à une variable explicative, soit par le *Mean Decrease Accuracy* (MDA), qui correspond quand à lui à l'apport d'une variable en terme de précision du modèle.

#### IMPORTANCE D'UNE VARIABLE PAR *Mean Decrease Accuracy*

Une mesure d'importance d'une variable donnée utilisée dans le *Random Forest* est l'accroissement moyen de l'erreur d'un arbre de la forêt lorsque les valeurs observées de cette variable sont permutées au hasard dans les échantillons OOB. [9]

Considérons  $B_t$  un échantillon bootstrap, utilisé pour la construction de l'arbre  $t$ , et  $OOB_t$  l'échantillon *Out Of Bag* associé, c'est-à-dire l'ensemble des observations qui ne sont pas utilisées pour la construction de l'arbre  $t$ . Le calcul de l'importance de la variable  $X_j$  est effectué par les étapes suivantes :

1. L'erreur commise sur  $OOB_t$  par l'arbre  $t$  (construit sur  $B_t$ ), notée  $erreur_{OOB_t}$  et correspondant à l'erreur quadratique moyenne en régression, est calculée pour chaque arbre  $t \in \{1, \dots, T\}$ , où  $T$  est le nombre d'arbres dans la forêt.
2. Les valeurs de la variable  $X_j$  sont permutées aléatoirement dans chaque  $OOB_t$ , ce qui donne un échantillon perturbé  $\widetilde{OOB}_t^j$ .
3. L'erreur commise sur  $\widetilde{OOB}_t^j$  par l'arbre  $t$  (notée  $erreur_{\widetilde{OOB}_t^j}$ ) est alors calculée pour chaque arbre  $t \in \{1, \dots, T\}$ .
4. L'importance de la variable  $X_j$ ,  $Imp_{MDA}(X_j)$ , est définie par la différence entre l'erreur moyenne d'un arbre sur l'échantillon OOB perturbé et celle sur l'échantillon OOB :

$$Imp_{MDA}(X_j) = \frac{1}{T} \sum_t^T (erreur_{\widetilde{OOB}_t^j} - erreur_{OOB_t})$$

Avec  $T$  le nombre total d'arbres dans la forêt.

Autrement dit, moins la permutation aléatoire de la variable  $X_j$  n'a d'incidence sur l'erreur commise par les arbres, moins cette variable est importante dans la prédiction.

### IMPORTANTANCE D'UNE VARIABLE PAR *Mean Decrease Impurity* [3]

Une deuxième mesure de l'importance d'une variable au sein d'une forêt aléatoire existe, *Mean Decrease Impurity*, qui se base quand à elle sur l'impureté des noeuds construits dans les arbres de la forêt.

Dans le cadre d'une régression, l'impureté d'un noeud est vue comme la variance de la variable cible  $Y$  dans ce noeud. L'importance d'une variable  $X_j$  pour la prédiction de la variable  $Y$  dans un arbre est alors vue comme la somme, pour tous les noeuds où  $X_j$  est utilisée, des diminutions d'impureté pondérées par la proportion d'individus atteignant ces noeuds. En reprenant les notations de "*Understanding variable importances in forests of randomized trees Supplementary materials*" [10].

$$\text{IMP}_{\text{MDI}}(X_j)_{\text{arbre } t} = \sum_{n \in t: \nu(s_n) = X_j} p(n) \times \Delta i(s_n, n)$$

avec

- $n$  les noeuds de l'arbre  $t$ .
- $s_n$  la coupure faite au noeud  $n$ .
- $p(n)$  la proportion des individus atteignant le noeud  $n$ .
- $\nu(s_n)$  la variable utilisée lors de la coupure  $s_n$ .
- Et,

$$\Delta i(s_n, n) = i(n) - p_L \times i(n_L) - p_R \times i(n_R)$$

avec

- $i(n)$  l'impureté du noeud  $n$ , s'agissant de la variance de  $Y$  dans le cadre de la régression.
- $p_L, p_R$ , respectivement la proportion des individus empruntant la branche gauche (L) ou droite (R) partant du noeud  $n$ .
- $i(n_R), i(n_L)$  l'impureté (ici la variance de  $Y$ ) des noeuds sortants (respectivement droit et gauche).

Enfin, l'importance de la variable  $X_j$  dans la forêt aléatoire est calculée comme la moyenne des importances par *Mean Decrease Impurity* à chaque arbre.

$$\text{IMP}_{\text{MDI}}(X_j) = \frac{1}{T} \sum_t^T \sum_{n \in t: \nu(s_n) = X_j} p(n) \times \Delta i(s_n, n)$$

L'importance des différents facteurs de risque obtenue par *Mean Decrease Impurity* pour la prédiction du *Best Estimate* stressé de la LoB unique d'assurance vie par le modèle de forêt aléatoire convenablement calibré est présenté dans la figure ci-dessous.

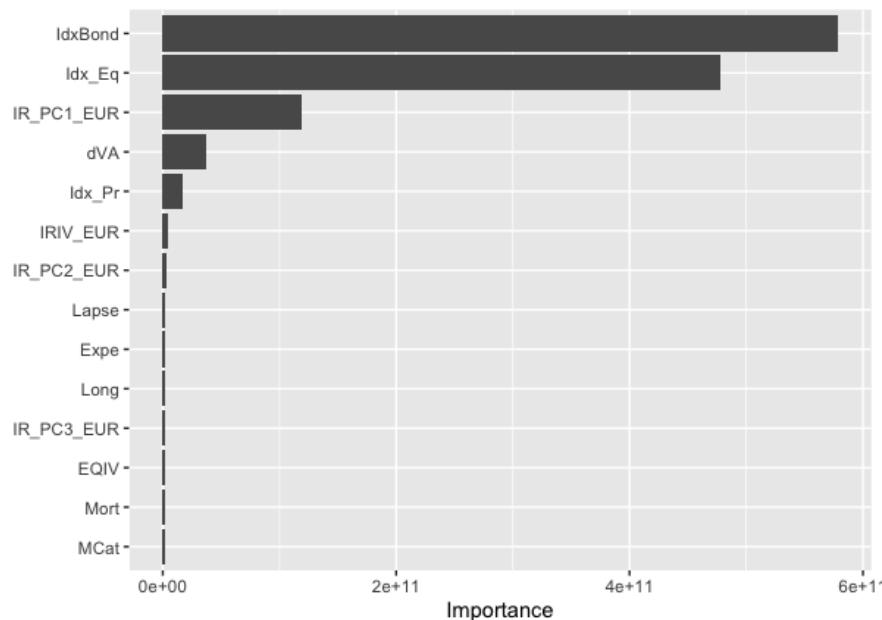


FIGURE 3.5 – Importance des variables dans le modèle de forêt aléatoire dans l'explication de BE de la LoB unique.

Nous pouvons remarquer que les variables `IdxBond`, `Idx_Eq`, `IR_PC1_EUR`, `dVA` et `Idx_Pr` ont un pouvoir explicatif très supérieur aux autres variables du modèle. Cela rejoint l'observation faite lors de la section d'analyse univariée, ce qui conforte la pertinence du modèle construit. Nous gardons les variables à faible pouvoir prédictif dans le modèle afin que la valeur du BE soit sensible à un stress sur l'un de ces facteurs de risque.

### La forêt aléatoire en pratique

Sur le logiciel R, la forêt aléatoire est implémentée dans plusieurs packages, le plus connu étant *RandomForest*. La fonction utilisée dans cette étude est *ranger()*, issue du package *Ranger*, qui est une version beaucoup plus rapide que *RandomForest*. Les principaux paramètres à optimiser lors de la construction d'une forêt aléatoire avec cette fonction sont les suivants :

- **num.trees** : le nombre d'arbres construits dans la forêt.
- **mtry** : le nombre de variables explicatives choisi aléatoirement pour la construction de chaque arbre.

- **min.node.size** : la taille minimal d'un noeud, c'est un critère d'arrêt.
- **max.depth** : la profondeur maximal d'un arbre, c'est un critère d'arrêt.

### 3.2.2 *EXtreme Gradient BOOSTing* (XGBoost)

L'*Extreme Gradient Boosting* est un modèle qui combine les efforts de plusieurs modèles faibles afin d'obtenir un modèle solide, et chaque modèle faible supplémentaire permet de réduire l'erreur quadratique moyenne du modèle global. Dans le cas de cette étude, le modèle faible utilisé est l'arbre de décision.

L'algorithme traverse des cycles qui construisent à plusieurs reprises de nouveaux modèles dit «faibles» et les combine en un modèle d'ensemble. Il commence le cycle en prenant un modèle existant dit «naïf», dans notre cas un arbre de décision, et en calculant les erreurs pour chaque observation dans l'ensemble de données. Il construit ensuite un nouveau modèle afin de prédire ces erreurs. Il ajoute les prédictions de ce modèle de prévision d'erreurs au modèle global. Ainsi de suite, de nouveaux modèles vont être agrégés afin d'expliquer un peu plus l'erreur à chaque itération.

#### La descente du gradient

Le gradient d'une fonction est un vecteur composé des dérivées partielles de cette fonction. Il se note ainsi :

$$\nabla f(x, y) = \left( \frac{\partial f}{\partial x}; \frac{\partial f}{\partial y} \right)$$

La descente du gradient est un algorithme d'optimisation qui permet de trouver le minimum d'une fonction convexe. Nous rappelons qu'une fonction  $f$  est dite convexe si pour tous points  $a$  et  $b$  de son domaine de définition, le segment reliant  $f(a)$  et  $f(b)$  est strictement au dessus du graphe de la fonction.

Le principe de la descente du gradient est le suivant [22] : on part d'un point aléatoire de la fonction et on cherche de quel côté il faut se déplacer pour atteindre le minimum de la fonction. Pour cela il suffit de trouver la pente négative la plus importante autour du point choisi. On "descend" donc dans cette direction et on réitère l'opération jusqu'à arriver au point le plus bas.

D'un point de vue mathématique voilà comment cela se traduit : On se donne un point initial  $x_0$  appartenant au domaine de définition de la fonction à minimiser, un seuil de tolérance  $\epsilon > 0$  et un pas  $\eta$ . Pour  $k = 0, 1, 2, \dots$  on effectue les étapes suivantes :

1. On calcule la valeur de la pente en ce point, c'est à dire le gradient.
2. Si la pente est plus faible que le seuil choisi :

$$\nabla f(x_k) \leq \epsilon$$

On arrête.

Sinon, on se déplace d'un certain pas  $\eta$  dans la direction de la pente qui descend :

$$x_{k+1} = x_k - \eta \times \nabla f(x_k)$$

3. On continue jusqu'à ce que le test d'arrêt soit vérifié.

### Le gradient *boosting*

Pour l'algorithme de gradient *boosting*, la fonction à minimiser est une fonction de perte entre la valeur réelle  $y$  et la valeur prédite  $\hat{y} = F_m(X)$ , avec  $F_m$  la fonction de prédiction. Voici ce qu'il se passe lorsque nous ajoutons un *modèle faible*  $\Delta_m(X)$  au modèle global : [19]

$$F_{m+1}(X) = F_m(X) + \eta \times \Delta_m(X)$$

Le résidu  $y - \hat{y}$  est un vecteur de direction. De plus, il pointe vers une meilleure prédiction et donc une perte plus faible entre les valeurs réelles et celles prédites. Cela suggère que le vecteur direction est également un gradient de fonction de perte. Ainsi on obtient :

$$\hat{y}_m = \hat{y}_{m-1} + \eta \times (-\nabla L(y, \hat{y}_{m-1}))$$

En comparant cette formule avec la formule du nouvel individu dans la descente du gradient, on constate que le gradient *boosting* se rapproche de la descente du gradient.

### Importance des variables

Tous comme avec la méthode *Random Forest*, il existe plusieurs méthodes permettant de relater de l'importance des variables au sein du modèle XGBoost.

La première approche consiste à compter le nombre de fois où une variable est utilisée pour séparer les données au sein de l'ensemble des arbres. La deuxième se base également sur le nombre de fois où une variable est utilisée dans un noeud, en pondérant cela par le nombre d'individus atteignant ce noeud. Enfin la troisième méthode de calcul d'importance de variables est la moyenne de la réduction de la fonction de coût quand une variable est utilisée pour la division d'un noeud.

La figure ci-dessous représente l'importance des variables sur la régression du BE stressé de la LoB unique d'assurance vie par XGBoost. Il est possible de constater que l'ordre d'importance des variables est sensiblement le même pour le modèle XGBoost que pour la forêt aléatoire.



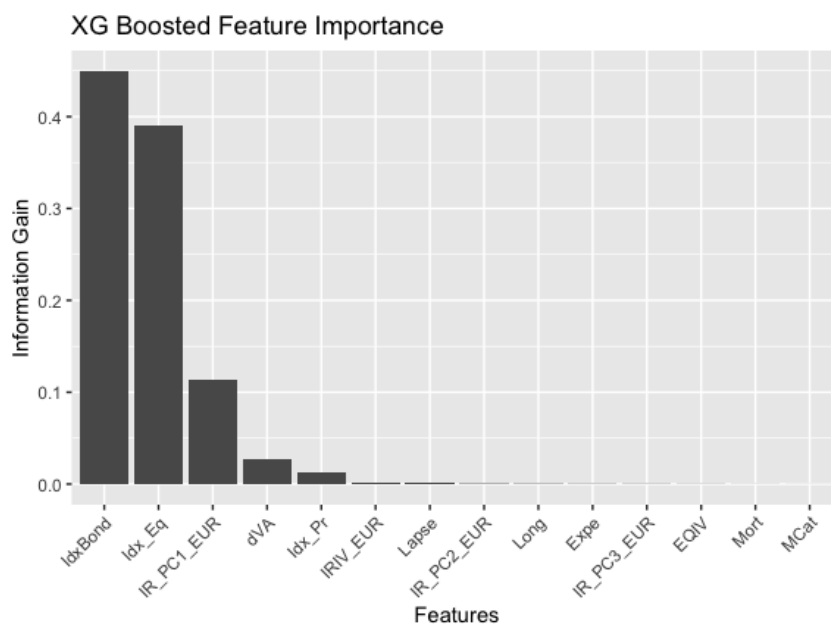


FIGURE 3.6 – Importance des variables dans le modèle XGBoost dans la régression du BE de la LoB unique.

### XGBoost en pratique

Lors de cette étude, le package R *xgboost*, et la fonction *xgb()* ont été utilisés.

Les principaux paramètres à optimiser lors de l'implémentation de la fonction *xgb()* sont les suivants :

- **nrounds** : le nombre d'itérations que fera le modèle, soit le nombre de modèles *faibles* agrégés.
- **eta** : la vitesse de convergence de la descente de gradient.
- **min\_child\_weight** : la taille minimale d'un noeud, c'est un critère d'arrêt.
- **max.depth** : la profondeur maximale d'un arbre, c'est un critère d'arrêt.
- **colsample\_bytree** : la part des variables explicatives présent en compte aléatoirement pour la construction de chaque arbre. Ce paramètre correspond au *mtry* dans la forêt aléatoire.

Ils ont été déterminés par validation croisée en utilisant la fonction de *tuning tune()* du package R *Caret*.

### 3.2.3 *Multivariate Adaptive Regression Splines* (MARS)

La description du modèle MARS sera inspirée de la partie 4 du chapitre 9 de *The Element of Statistical Learning, Data Mining, Inference and Prediction* [12].

La méthode de splines<sup>4</sup> de régression adaptative multivariée, appelée MARS (*Multivariate Adaptive Regression Splines*), fournit une approche pratique pour capter les non-linéarités entre les variables explicatives en utilisant des points de coupure, appelés noeuds ou *knots*. Le modèle MARS utilise des splines linéaire, autrement dit, une régression linéaire différente entre chaque point de coupure est ensuite calibrée afin de coller au mieux aux données. Plus il y a de points de coupure, plus la régression sera proche des données d'apprentissage, mais cela peut également amener à un sur-apprentissage. Pour éviter le sur-apprentissage, une étape d'élagage, de *pruning*, a ensuite lieu afin de retirer les points de coupure qui n'apportent pas de grande amélioration au modèle.

#### Principe général

MARS utilise des fonctions de base linéaires par morceaux de la forme  $(x - t)_+$  et  $(t - x)_+$ , avec  $( )_+ : x \mapsto \begin{cases} 0, & \text{si } x < 0 \\ x, & \text{sinon} \end{cases}$ .

Chaque fonction de base est donc une fonction linéaire par morceaux en fonction de noeuds  $t$ , ce sont des splines linéaires. L'idée est de construire des paires de fonctions de base pour chaque variable explicative, avec des noeuds à chaque point de donnée d'apprentissage  $x_{i,j}$ . L'ensemble des fonctions de base considérées est alors

$$\mathcal{B} = \{(X_j - t)_+, (t - X_j)_+\}_{t \in x_{1,j}, x_{2,j}, \dots, x_{N,j} \quad j=1,2,\dots,p.}$$

Le modèle MARS prend alors la forme

$$f(X) = \beta_0 + \sum_{m=1}^M \beta_m h_m(X) \tag{3.1}$$

avec  $h_m(X)$  une fonction de  $\mathcal{B}$  ou un produit de plusieurs fonctions de  $\mathcal{B}$ ,  $M$  le nombre maximal de termes considérés dans le modèle et les coefficients  $\beta_m$  sont estimés par minimisation de l'erreur quadratique, c'est à dire par régression linéaire standard.

#### Construction du modèle MARS : Sélection *forward* puis élimination *backward*

Lors de sa construction, le modèle MARS va considérer toutes les variables explicatives et chacun des points comme noeud potentiel à travers une procédure adaptative en deux étapes : une étape de sélection *forward* (vers l'avant) des termes, puis une étape d'élimination des termes *backward* (vers l'arrière).

---

4. Une spline est une fonction définie par morceaux par des polynômes. Soit une série de points appelés noeuds, la spline est la fonction définie par morceaux constituée d'un polynôme différent sur chaque intervalle entre deux noeuds. Le degré de la spline est défini comme étant le degré maximal des polynômes utilisés.

*Étape de sélection forward :*

Lors de l'étape de sélection des termes, le choix des fonctions  $h_m$  est fait de la manière suivante ; Le modèle commence par considérer seulement la fonction constante  $h_0(X) = 1$ , et toutes les fonctions de  $\mathcal{B}$  sont candidates pour être rajoutées au modèle. À chaque étape, l'algorithme va considérer comme nouvelle paire de fonctions de base tous les produits d'une fonction  $h_m(X)$  présente dans le modèle  $\mathcal{M}$  et d'une paire de fonctions de  $\mathcal{B}$ . C'est à dire que le terme rajouté au modèle sera de la forme

$$\beta_{m+1}h_l(X) \times (X_j - t)_+ + \beta_{m+2}h_l(X) \times (t - X_j)_+, \quad h_l(X) \in \mathcal{M},$$

de manière à ce qu'il produise la plus grande diminution de la somme des erreurs quadratiques résiduelles. Les termes sont ainsi rajoutés un à un de manière itérative jusqu'à ce que le nombre de termes maximal, spécifié par l'utilisateur, soit atteint ou qu'aucun nouveau terme n'améliore le modèle.

Il existe une restriction dans la formation des termes du modèle : chaque variable explicative ne peut apparaître qu'une seule fois dans un produit. Cette restriction empêche la formation de degré supérieur d'un prédicteur, qui pourrait varier trop fortement aux extrémités des données.

La figure ci-dessous, extraite du Chapitre 7 de *Hands-On Machine Learning with R*, B. BOEHMKE & B. GREENWELL [11] permet de visualiser comment fonctionne la première phase de construction du modèle MARS en dimension 2.

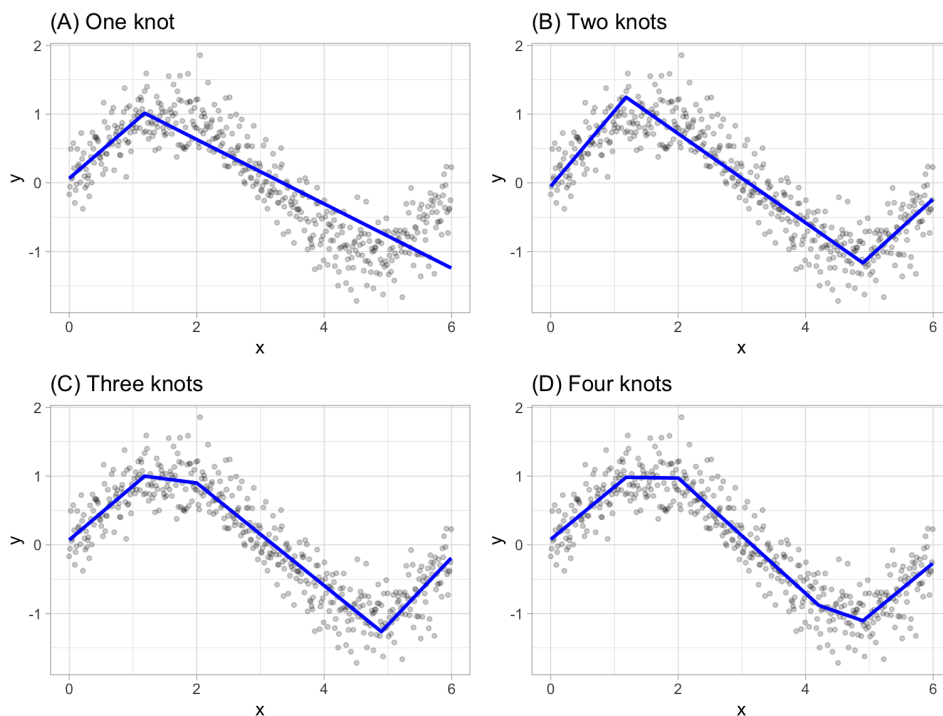


FIGURE 3.7 – Schéma du fonctionnement du modèle MARS

*Étape d'élimination backward* :[15]

Une fois ce processus de sélection *forward* terminé, le modèle construit peut avoir tendance à sur-apprendre les données d'apprentissage. L'algorithme passe alors à une étape d'élagage par une procédure itérative *backward* d'élimination de termes. À chaque itération, la procédure va retirer du modèle le terme qui contribue le moins au modèle, afin d'obtenir le meilleur modèle estimé  $\hat{f}_\lambda$  de taille (nombre de termes)  $\lambda$ .

Le critère utilisé dans ce processus *backward* pour déterminer quel termes apporte le moins d'information au modèle est un GCV (*Generalized Cross Validation*) modifié. Le critère GVC part de l'idée de minimisation de l'erreur quadratique moyenne du modèle. Il est donné par :

$$GCV(M) = \frac{1}{N} \sum_{i=1}^N \left[ \frac{y_i - \hat{f}_M(X_i)}{1 - \frac{P(M)^*}{N}} \right]^2$$

avec  $y_i$  les réponses observées,  $\hat{f}_M(X_i)$  les valeurs prédites par le modèle,  $N$  est le nombre d'observations et  $M$  représente le nombre maximal de termes dans le modèle.  $P(M)$  représente le nombre effectif de paramètres ce qui est une mesure de pénalité pour la complexité. Il est généralement calculé par :

$$P(M) = \text{trace}(B(B^T B)^{-1} B^T) + 1$$

avec  $B$  la matrice des fonctions de bases (c'est à dire des termes compris dans le modèle) de taille  $M \times N$ .

Une forme modifiée  $P^*$  est utilisé dans MARS, qui est obtenue par

$$P^*(M) = P(M) + dM$$

avec  $M$  le nombre total de termes dans le modèle et  $d$  qui vaut 2 lorsqu'aucune interaction entre variables n'est autorisée ou 3 sinon.

### MARS en pratique

Le modèle de spline multivariée utilisé au cours de cette étude est implémenté dans le package *earth* de R, via la fonction *earth()*. Voir *Notes on Earth package* [18].

Les paramètres à choisir par l'utilisateur, et donc à optimiser afin que la fonction se rapproche au mieux des données sont les suivants :

- **degree** : le degré maximal d'interaction entre les différents termes.
- **nk** : le nombre maximal de termes envisagés avant l'étape de d'élagage.
- **nprune** : le nombre maximal de termes dans le modèle final.
- **endspan** : le nombre minimal de points avant le premier noeud et après le dernier noeud.

Le degré maximal d'interaction (degree) et le nombre de termes maximal dans le modèle final (nprune) ont été déterminés par validation croisée à l'aide de la fonction *train()* du package R *Caret*. Les autres hyper-paramètres ont été optimisés, par la suite, par une validation croisée 5-fold.

### 3.2.4 *Support Vector Machine* : SVM

L'algorithme de machines à vecteurs de support ou séparateur à vastes marges (SVM, *Support Vector Machine*, en anglais) est initialement conçu pour des problèmes de classification. Le principe de cet algorithme est de trouver un hyperplan optimal permettant de séparer les données en des classes disjointes. Il existe désormais une extension de cette méthode pour résoudre des problèmes de régression, parfois appelée Support Vector Regression (SVR), ce qui sera le cas dans la suite.

L'algorithme classique de SVM dans le cadre de la classification ne sera pas explicité ici, nous entrons directement dans le cadre de la régression.

#### Principe général du SVM pour de la régression : SVR

Cette sous section sera inspirée de la section 2.4 de [21], et de [6].

Contrairement au SVM classique qui va chercher l'hyperplan séparant au mieux les données en maximisant la marge d'écart entre l'hyperplan et les vecteurs supports (correspondant aux points les plus proches de la limite), le SVR va chercher l'hyperplan permettant de se rapprocher au mieux des données. Pour cela, il va chercher à résoudre le problème d'optimisation suivant :

*Recherche de  $(\omega, \epsilon, \xi)$  qui minimise*

$$\frac{1}{2} \|\omega^2\| + C \sum_{i=1}^n |\xi_i| \text{ sous la contrainte}$$

$$|y_i - \omega_i x_i| \leq \epsilon + \xi_i, \quad i = 1, \dots, n$$

Où  $\omega$  correspond à la distance entre un point et l'hyperplan de régression,  $\epsilon$  correspond à la marge d'erreur acceptée,  $\xi$  représente la distance entre un point hors de la marge de risque et la marge de risque, permettant d'autoriser certains points *ressorts* en dehors de la marge, et  $C$  est un paramètre qui pénalise les termes qui sont en dehors de la marge d'erreur  $\epsilon$ . Il est possible de visualiser cela sur le schéma présenté en figure 3.9 ci dessous, prenant l'exemple d'un cas simple.

Dans cet exemple, le couloir entre les deux droites noires représente la marge d'erreur et la droite rouge correspond à la droite de régression solution du problème d'optimisation ci-dessus.

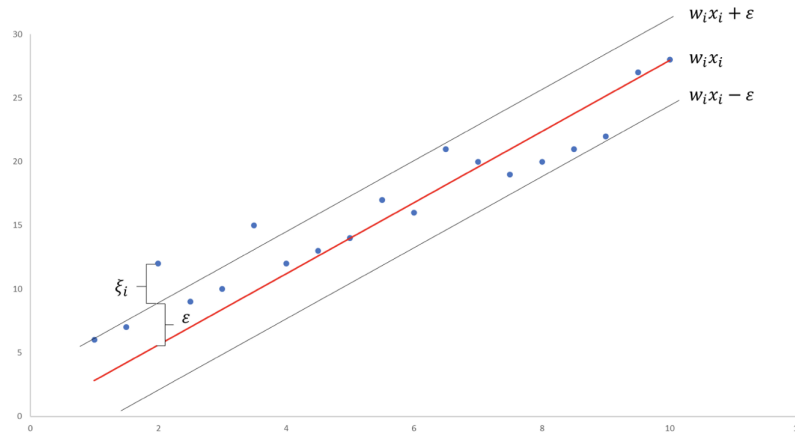


FIGURE 3.8 – Exemple illustratif d'un cas simple de SVR

(Source : <https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2>)

Afin de capter les non-linéarités présentes dans les données, l'astuce des noyaux peut être appliquée. Cette astuce consiste à projeter les données dans un espace de Hilbert  $\mathcal{H}$  de dimension supérieure, appelé espace de représentation, dans lequel on pourrait relier les données de manière linéaire. En pratique, il n'est pas nécessaire de connaître  $\mathcal{H}$  mais seulement le noyau  $K$ .

**Définition :** (Aronszajn, 1950)

Une fonction  $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$  est un **noyau** si et seulement si elle est (symétrique) définie positive. C'est-à-dire ssi

1.  $K(x, x') = K(x', x) \forall (x, x') \in \mathcal{X}^2$
2.  $\forall (x_1, \dots, x_N) \in \mathcal{X}^N$  et  $\forall (a_1, \dots, a_N) \in \mathbb{R}^N$

$$\sum_{i=1}^N \sum_{j=1}^N a_i a_j K(x_i, x_j) \geq 0$$

Voici les principales fonctions de noyau utilisées dans le cadre d'un SVM :

- Le noyau linéaire :

$$\forall (x, x') \in \mathcal{X}^2, K(x, x') = \langle x, x' \rangle$$

- Le noyau polynômial, pour  $c \in \mathbb{R}, d \in \mathbb{R}$  :

$$\forall (x, x') \in \mathcal{X}^2, K(x, x') = (c + \langle x, x' \rangle)^d$$

- Le noyau radial, pour  $\sigma > 0$  :

$$\forall (x, x') \in \chi^2, K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

- Le noyau sigmoïde, pour  $c \in R, d \in R$  :

$$\forall (x, x') \in \chi^2, K(x, x') = \tanh(c \langle x, x' \rangle + d)$$

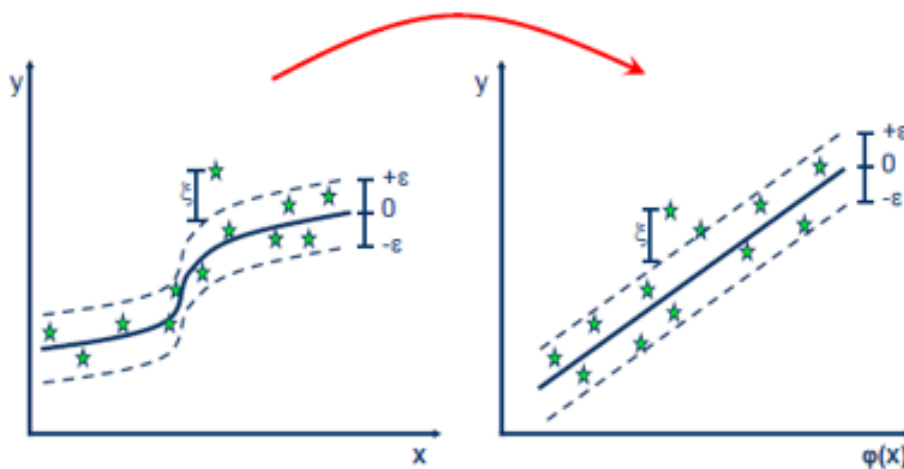


FIGURE 3.9 – Exemple illustratif de l'astuce du noyau lors d'un SVR

### SVM en pratique

La fonction `ksvm()` du package R `kernelab` a été utilisée pour modéliser la régression du *Best Estimate* en fonction des facteurs de risque par la méthode de *Support Vector Machine*.

Les hyper-paramètres optimisés pour obtenir le modèle le plus adéquat aux données ont été les suivants :

- **type** : le type de svm mis en place (régression, classification binaire, classification multi-classes, ...)
- **kernel** : la fonction de noyau utilisée.
- **C** : ajoute une pénalité pour chaque point mal classé.
- **gamma** : contrôle l'influence d'un unique point d'apprentissage.
- **epsilon** : seuil de tolérance.

### 3.3 Calibration des modèles et résultats obtenus

Le choix des hyper-paramètres des différents modèles a été fait par validation croisée, plus particulièrement par la méthode k-fold.

#### 3.3.1 Validation croisée par k-fold

Le processus du tuning des hyper-paramètres consiste à optimiser les paramètres du modèle de telle sorte que le modèle se rapproche le plus des données. Si le choix de ces paramètres est effectué directement sur la base d'apprentissage, le modèle risque de sur-apprendre. On parle de sur-apprentissage lorsqu'un modèle décrit extrêmement bien les données d'apprentissage mais que ses performances sur de nouvelles données non-explorées se détériorent.

L'utilisation de la validation croisée permet d'éviter le sur-apprentissage, en testant les performances du modèle sur différentes sous-parties des données de façon itérative. Une des approches de validation croisée s'appelle la méthode des k-fold. Cette méthode est celle utilisée au cours de ce mémoire pour effectuer le choix des hyper-paramètres optimaux. Elle sera présentée en suivant le schéma de la figure 3.6 ci-dessous.



FIGURE 3.10 – Schématisation de la validation croisée par k-fold  
 Source : Wikipedia ([https://en.wikipedia.org/wiki/Cross-validation\\_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics)))

La validation croisée par k-fold divise les données en k sous-parties, puis effectue les itérations k suivantes :

- Itération 1 : Le modèle est construit à partir des (k-1) dernières sous-parties, puis ses performances sont testées sur la première sous-partie, considérée comme la base de test pour cette itération.
- Itération 2 : La base de test est désormais la deuxième sous-partie des données et le modèle est construit à partir du reste des données.



- ...
- Itération k : La base de test est désormais la k-ième sous-partie, et le modèle est construit à partir des (k-1) autres sous-parties des données.

Une fois ces k itérations effectuées, les performances du modèle sont obtenues en faisant la moyenne des performances calculées à chaque itération sur les différentes bases de test.

### 3.3.2 Modèles finaux et résultats graphiques

Voici les paramètres finaux utilisés pour la construction des modèles et les résultats de ces modèles ainsi construits :

#### Random Forest

- mtry = 8
- ntree = 900
- node\_size = 3

La forêt aléatoire finale a été construite à partir de la fonction *ranger()* pour ces valeurs de paramètres, sur l'ensemble de la base d'apprentissage. Ces valeurs d'hyperparamètres ont été obtenues par la validation croisée présentée précédemment. Autrement dit, la forêt aléatoire finale est composée de 900 arbres décisionnels indépendants, chacun construits à partir de 8 variables explicatives aléatoirement choisies. De plus elle a été construite à partir du critère d'arrêt suivant : 3 individus minimum par noeud.

Avec ce modèle, voici les résultats obtenus sur la base de validation (nous rappelons que les résultats présentés au long du rapport sont ceux obtenus sur la LoB unique d'assurance vie) :

- Erreur relative absolue moyenne = 0,36%
- Erreur relative absolue maximum = 1,4%
- Racine de l'erreur quadratique moyenne (RMSE)<sup>5</sup> = 311

D'un point de vue visuel, il est possible de se rendre compte de la précision du modèle grâce au graphique suivant.

---

5. La RMSE sera définie dans la sous-section suivante, elle est à minimiser.

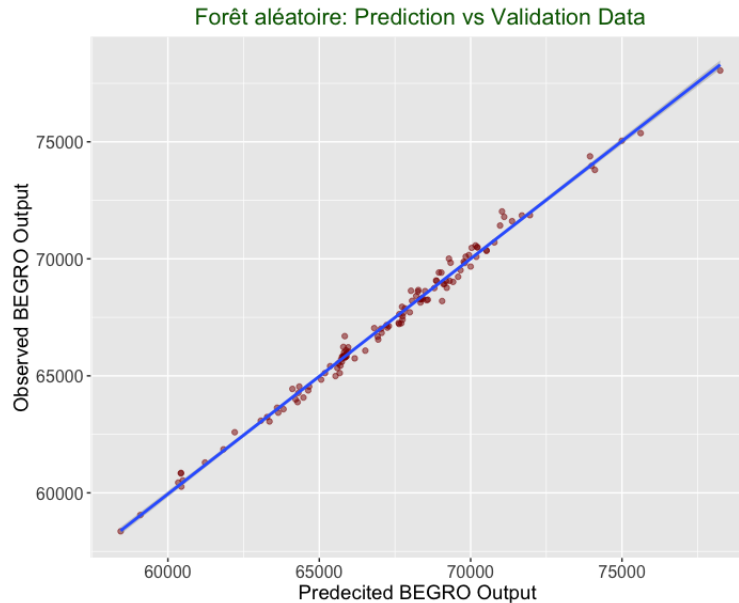


FIGURE 3.11 – Valeurs prédites en fonction des valeurs de validation réelles sur la LoB unique par facteur de risque

Le graphique ci-dessus représente les valeurs de *Best Estimate* prédites par la forêt aléatoire ainsi construite pour les 125 points de la base de validation, en fonction des valeurs réelles de la base de validation. En bleu est tracée la droite d'équation  $y = x$ . Plus les points sont proches de cette droite, et plus la prédiction est exacte.

### XGBoost

Le modèle XGBoost final est construit à partir des paramètres suivants, choisis par la validation croisée à l'aide de la fonction *train()* du package R *Caret* :

- nrounds = 1900
- max.depth = 4
- colsample\_bytree = 0,6
- eta = 0,05
- min\_child\_weight = 200

À partir de ce modèle XGBoost, les prédictions faites sur les 125 valeurs de BE de la base de validation sont présentées sur la figure suivante, et les résultats obtenus par ce modèle sur la base de validation sont les suivants :

- Erreur relative absolue moyenne = 0,29%

- Erreur relative absolue maximum = 1,3%
- Racine de l'erreur quadratique moyenne (RMSE) = 268

Ici encore, le graphique représente en abscisse les valeurs de *Best Estimate* prédites par le modèle et en ordonnée les valeurs observées dans la base de validation, avec en bleu la droite d'équation  $y = x$  reflétant une prédiction exacte. Nous pouvons remarquer que les points prédits par XGBoost sont en majorité légèrement plus proche de la droite bleue par rapport aux points prédits par la forêt aléatoire. Néanmoins, il est également observable que certains points s'écartent plus fortement de l'axe bleu. Cela reflète bien les résultats donnés par les trois métriques d'erreur explicitées ici.

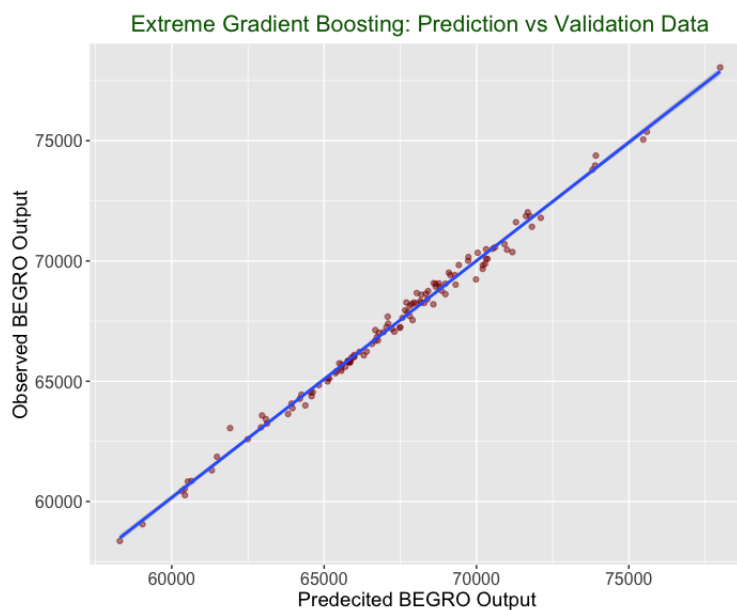


FIGURE 3.12 – Valeurs prédites en fonction des valeurs de validation réelles par XGB

### Multivariate Adaptive Regression Spline

Le modèle MARS a été calibré par la fonction `earth()` de R, avec les paramètres suivants :

- `degree = 4`
- `nk = 150`
- `endspan = 1300`
- `fast.k = 16`

C'est à dire que le degré maximal d'interaction entre les variables explicatives est de 4, le nombre maximal de termes considérés lors de l'étape forward est 150. Le paramètre *endspan* permet de n'autoriser la création d'un noeud qu'à partir de 1300 points des données. Cela évite la prise un compte d'un effet de bord non représentatif, et ainsi d'être plus généralisable au bord de l'intervalle d'apprentissage.

Les prédictions du modèle MARS sont représentées graphiquement par rapport aux valeurs réelles de *Best Estimate* de la base de validation sur la figure suivante. Et ce modèle conduit aux résultats suivants :

- Erreur relative absolue moyenne = 0,17%
- Erreur relative absolue maximum = 0,65%
- Racine de l'erreur quadratique moyenne (RMSE) = 147

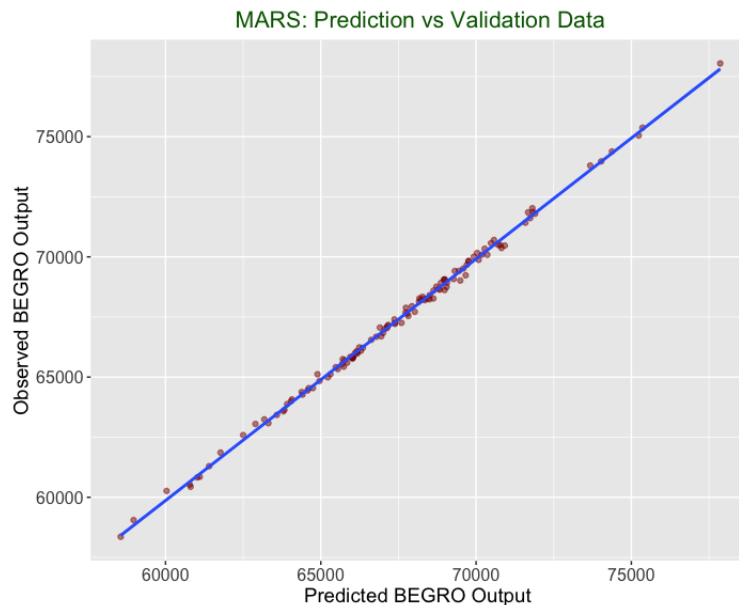


FIGURE 3.13 – Valeurs prédites en fonction des valeurs de validation réelles par MARS

### Support Vector Machine Regression

Enfin, le modèle final de SVR a été construit avec les hyper-paramètres ci-dessous :

- $\text{tol} = 0,1$
- $\text{nu} = 0,16$
- $\text{epsilon} = 0,6$
- $\text{type} = \text{"nu-svr"}$

Ainsi calibré, le modèle obtient les résultats suivants sur la base de validation (pour la Lob unique d'assurance vie) :

- Erreur relative absolue moyenne = 0,24%
- Erreur relative absolue maximum = 0,85%
- Racine de l'erreur quadratique moyenne (RMSE) = 208

Les prédictions du modèle SVR sont visualisables sur le graphique ci-dessous, représentant les valeurs réelles en fonction des valeurs prédites par le modèle SVR sur la base de validation. Les points sont ici encore plus proches de l'axe  $y=x$  représenté en bleu que les prédictions de Forêt aléatoire et de XGBoost, ce qui dénote d'une meilleure précision, mais légèrement moins que MARS ; comme il ressort des métriques d'erreurs.

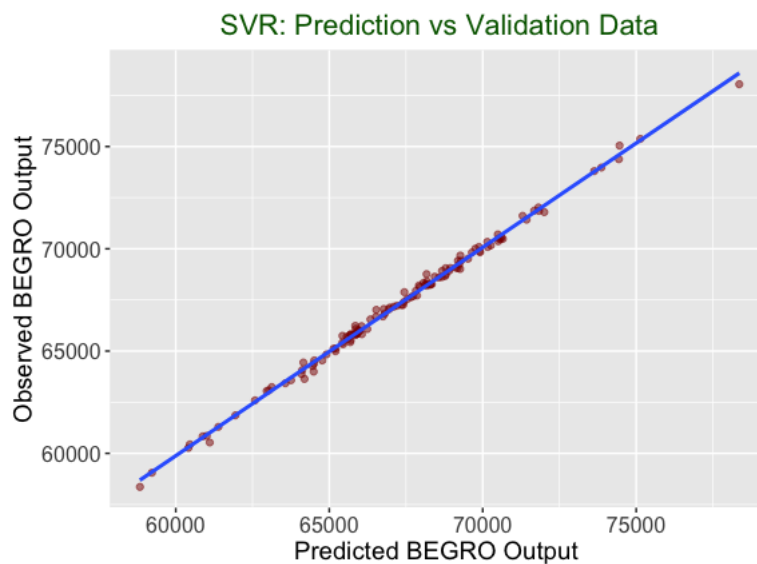


FIGURE 3.14 – Valeurs prédites en fonction des valeurs de validation réelles par SVR

## 3.4 Comparaison et choix du meilleur modèle

Afin de comparer les modèles entre eux, il est nécessaire de déterminer certaines métriques à appliquer sur les résultats de chacun des modèles sur base de validation. Les métriques suivantes ont été choisies pour comparer les modèles entre eux. Il faut savoir qu'il existe de nombreuses autres métriques notamment certaines permettant de tenir compte de la complexité des modèles.

### 3.4.1 Métriques utilisées pour la comparaison

Afin de comparer la performance des différents modèles entre eux, nous avons eu recours aux différentes métriques suivantes.

### La RMSE

La RMSE, *Root Mean Square Error* en anglais, correspond à la racine carré de l'erreur quadratique moyenne. Cette métrique est calculée de la manière suivante :

$$RMSE = \sqrt{\frac{\sum (y_{pred} - y)^2}{N}}$$

Avec,

- $y$  le *Best Estimate* stressé de la base de validation.
- $y_{pred}$  le *Best Estimate* prédit par le modèle à partir des variables explicatives de la base de validation.
- $N$  le nombre de termes dans la base de validation.

### La WRE

La WRE, *Weighed Relative Error*, correspond à l'erreur relative moyenne pondérée par la valeur de la variable cible. Elle s'obtient par le calcul suivant :

$$WRE = \frac{|y| \times |erreur\ relative|}{\sum |y|}$$

avec

$$erreur\ relative = \frac{y_{pred} - y}{y}$$

### Le $R^2$

Le  $R^2$  correspond au pourcentage de la variance expliquée par le modèle. Il se calcule par la formule qui suit :

$$R^2 = 1 - \frac{\sum (y_{pred} - y)^2}{\sum (y - y_{moy})^2}$$

avec  $y_{moy}$  la moyenne des *Best Estimate* sur la base de validation.

### 3.4.2 Résultats obtenus et comparaison des modèles

Voici un récapitulatif des métriques calculées sur les prédictions de la base de validation en fonction du modèle de *Machine Learning* appliqué.

Modèles	RMSE	WRE	$\epsilon \max^6$	$R^2$
<b>Polynôme actuel</b>	<b>150.5</b>	<b>0.19%</b>	<b>0.5%</b>	<b>99.8%</b>
<b>Random Forest</b>	311	0.36%	1.4%	99.1%
<b>XGBoost</b>	268	0.30%	1.3%	99.4%
<b>SVM</b>	208	0.24%	0.85%	99.6%
<b>MARS</b>	<b>146.7</b>	<b>0.173%</b>	<b>0.66%</b>	<b>99.8%</b>

FIGURE 3.15 – Métriques calculées sur la base de validation pour chaque modèle.

Comme il était possible de le remarquer sur les graphiques précédents représentant les valeurs prédites en fonction des valeurs réelles des points *Best Estimate* de la base de validation, le modèle MARS est le plus précis des modèles de machine learning testés sur tous les critères considérés. Il est également plus précis au regard du RMSE, de la WRE et du  $R^2$  que la méthode polynomiale actuellement en place dans le modèle interne. Néanmoins, il est à noter que bien que meilleur au regard de la moyenne, il commet des erreurs maximales plus élevées que le polynôme actuel. En effet, 4 points de la base de validation sont prédits avec une erreur relative absolue supérieure à 0.5%.

Il est également intéressant de regarder le temps qu'a nécessité la mise en place de ces modèles. Pour cela, le tableau suivant va comparer les temps de 3 étapes de modélisations :

- Le temps de tuning, *Tuning Time*, correspondant au temps qu'il faut pour trouver les hyper-paramètres optimaux des modèles en fonction des données. Par exemple le nombre d'arbres de la forêt, le degré d'interaction dans MARS, etc.
- Le temps de fitting, *Fitting Time*, qui correspond quant à lui au temps de fitting de la fonction paramétrique, en sachant quels hyper-paramètres utiliser. Par exemple, le temps de construction de la forêt aléatoire en utilisant un nombre d'arbres prédéfini, un nombre de variables explicatives à utiliser par arbre prédéfini, etc.
- Et enfin le temps de prédiction, *Prediction Time*, correspondant au temps qu'il faut au modèle pour prédire les 125 points de validation.

	Random Forest	XGB	SVM	MARS
<b>Tuning Time</b>	<b>25min</b>	7-8 h	4h	<b>30min</b>
<b>Fitting Time</b>	2min	3min30	2min30	<b>52 sec</b>
<b>Prediction Time</b>	1.25 sec	0.115 sec	0.035 sec	<b>0.01 sec</b>

FIGURE 3.16 – Temps de construction et de prédiction des modèles

Encore une fois, le modèle MARS surpasse les trois autres modèles en termes de temps de *fitting* et de prédiction. Cependant, il est légèrement plus rapide de trouver les hyper-paramètres optimaux de la forêt aléatoire que ceux du modèle MARS, mais le

temps de prédiction nécessité par la forêt aléatoire est considérablement plus important. Or, cet écart de temps de prédiction peut poser problème lors du calcul de la distribution complète du *Best Estimate* stressé, par l'application de la forme paramétrique obtenus aux  $2^8$  scénarios de la *Risk Factor Scenario Set*.

### 3.4.3 Choix et étude du meilleur modèle

Comme les deux tableaux précédents le mettent en exergue, le modèle MARS est le plus performant des modèles testés dans cette étude, et ce sur toutes les métriques de comparaison utilisées. De plus, il est également le plus rapide en terme de *Fitting Time* et *Prediction Time*, et l'un des plus rapide en *Tunning Time*, juste derrière Random Forest.

Le modèle MARS ne nécessite aucune hypothèse sous-jacente, comme il est précisé dans *MARSpline*[1] il n'y en a donc aucune à vérifier, mais il reste néanmoins intéressant d'étudier les résidus du modèle, afin notamment de vérifier qu'il n'y a pas d'anomalie.

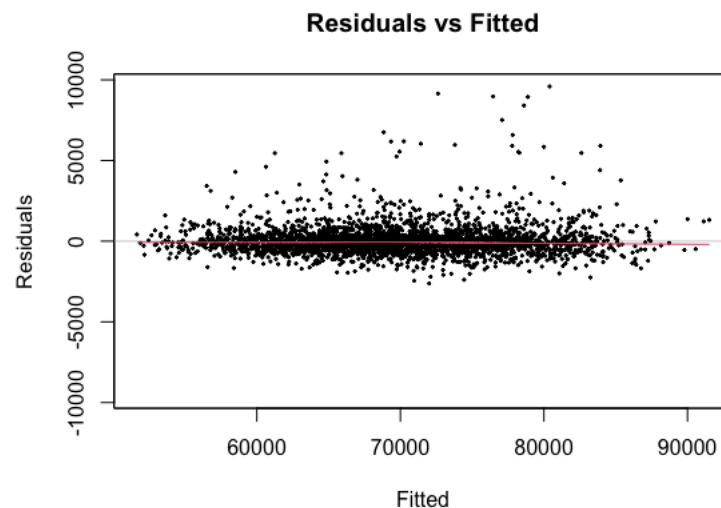


FIGURE 3.17 – Résidus en fonction de la valeur du *Best Estimate*.

Ce graphique représente les résidus du modèle en fonction de la valeur du *Best Estimate* stressé approché de la base d'apprentissage LSMC. Sur ce graphique, nous constatons que les résidus sont centrés sur 0. De plus, les résidus ont une variance plutôt constante en fonction de la valeur du *Best Estimate* approché. Ceci est une bonne chose puisque cela sous-entend que les performances du modèle ne sont pas meilleures pour certaines valeurs de *Best Estimate* et inférieures pour d'autres. Néanmoins, nous remarquons que certains résidus positifs sont très importants.



#### CONCLUSION DU CHAPITRE :

À partir d'une base de donnée contenant 25 000 valeurs imprécises de *Best Estimate* stressé en fonction de 14 facteurs de risque, 4 modèles de machine learning ont été calibrés : le RandomForest, l'Extrême Gradient Boosting, le Support Vector Machin adapté à la régression et le Multivariate Adaptive Regression Spline. Ces modèles ont été optimisés par validation croisée dans le but de choisir les hyper-paramètres de telle sorte que les modèles s'adaptent le mieux possible aux données.

Les performances des 4 modèles ainsi construits ont ensuite été calculées sur une base de validation, contenant 125 valeurs de Best Estimate fiables en fonction des 14 facteurs de risque. Afin d'obtenir des valeurs précises de *Best Estimate* pour chacun des 125 scénarios, 1000 simulations secondaires Monte Carlo ont été réalisées, amenant à un budget de 125 000 scénarios pour la construction de cette base de validation.

En comparant les performances des modèles entre eux ainsi qu'avec celles du polynôme actuellement utilisé au sein du modèle interne de Generali, il ressort que le modèle MARS surpasse les autres. De plus, ce modèle est également très intéressant en terme de temps de calcul. L'étude des résidus ne révèle aucune anomalie.

#### TRANSITION :

*Après avoir constaté que la fonction proxy du Best Estimate stressé obtenue par le modèle MARS surpasse légèrement le polynôme actuellement utilisé, et ce en un temps de calibration faible, il est intéressant de connaître l'impact qu'aurait un changement de fonction proxy du Best Estimate sur la valeur du SCR. Le chapitre suivant sera consacré au calcul et à l'analyse de cet impact sur le SCR.*



## Chapitre 4

# Impact d'un changement de forme paramétrique sur le SCR final

Dans ce chapitre encore, les impacts et les analyses seront faites seulement sur la LoB (Ligne d'activité) regroupant les produits d'assurance vie avec interaction ALM.

### 4.1 Différence entre la fonction polynomiale actuelle et la fonction obtenue par MARS

#### 4.1.1 Le polynôme actuel en détails

La fonction paramétrique actuelle permettant d'évaluer les provisions *Best Estimate* stressées de la LoB regroupant les produits d'assurance vie présentant des interactions ALM, obtenue par une régression polynomiale de degré 3 avec intégration de termes croisés allant jusqu'au degré 4 et application d'un algorithme *Adaptive Stepwise Forward* de sélection des termes, s'écrit sous la forme suivante :

$$\begin{aligned} f(RF) = & 66011 \\ & + 31346 \times IdxBond + 3165 \times Idx\_Eq \\ & - 863 \times IR\_PC1\_EUR - 35 \times dVA \\ & + 3377 \times Idx\_Pr + 6 \times IR\_PC1\_EUR \times dVA \\ & + 5899 \times Idx\_Eq \times IdxBond + 97 \times IdxBond \times dVA \\ & + 138520 \times IRIV + 0,11 \times dVA^2 \\ & + 275 \times Idx\_Eq^2 + 199 \times IR\_PC2\_EUR \\ & - 1860 \times Idx\_Eq^2 \times IdxBond \\ & + 55783 \times IdxBond^2 \\ & + 2691 \times IR\_PC1\_EUR \times IdxBond \\ & - 25595 \times Idx\_Eq \times IdxBond^2 \\ & + 4 \times Idx\_Eq \times dVA + 12 \times Idx\_Pr \times dVA \\ & + 37 \times Long \\ & - 36 \times Idx\_Eq \times IdxBond \times dVA \end{aligned}$$

+ ...

Seuls les 20 premiers termes de la forme paramétrique polynomiale sont explicités ci-dessus afin d'éviter trop de lourdeur. La fonction comporte en réalité 150 termes. De plus, les coefficients ont été arrondis à l'unité.

En prêtant attention aux 20 premiers termes explicités ci-dessus, il est remarquable que les facteurs de risque ressortant le plus souvent sont  $IdxBond$ ,  $Idx\_Eq$ ,  $dVA$  et  $IR\_PC1\_EUR$ .

#### 4.1.2 La fonction de régression calibrée par le méthode MARS en détails

La fonction du *Best Estimate* stressé obtenue grâce à la méthode MARS (*Multivariate Adaptive Regression Spline*) autorisant un degré d'interaction entre les termes allant jusqu'à 4, comporte quant à elle 109 termes. En explicitant seulement les premiers termes du modèle, voici l'allure de la fonction paramétrique obtenue par MARS<sup>1</sup> :

$$\begin{aligned}
 f(RF) &= 70156.110146 \\
 &+ 34192 \times h(IdxBond - 0.0958) \\
 &- 35602 \times h(0.0958 - IdxBond) \\
 &+ 3780 \times h(Idx\_Eq - 0.605563) \\
 &- 4987 \times h(0.605563 - Idx\_Eq) \\
 &- 517 \times h(IR\_PC1\_EUR + 0.337943) \\
 &+ 639 \times h(-0.337943 - IR\_PC1\_EUR) \\
 &- 16 \times h(dVA - 47.6094) \\
 &+ 30 \times h(47.6094 - dVA) \\
 &+ 4770 \times h(Idx\_Pr - 0.261052) \\
 &- 4478 \times h(0.261052 - IdxPr) \\
 &- 5 \times h(IR\_PC1\_EUR + 2.1965) * h(47.6094 - dVA) \\
 &- 12180 \times h(0.605563 - Idx\_Eq) * h(IdxBond + 0.017332) \\
 &+ 25177 \times h(0.605563 - Idx\_Eq) * h(-0.017332 - IdxBond) \\
 &- 27 \times h(IdxBond + 0.0283) * h(47.6094 - dVA) \\
 &+ 107 \times h(-0.0283 - IdxBond) * h(47.6094 - dVA) \\
 &- 4448 \times h(0.605563 - IdxEq) * h(-0.017332 - IdxBond) * IR\_PC1\_EUR \\
 &+ 122453 \times h(IRIV\_EUR + 0.001226) - 25727h(-0.001226 - IRIV\_EUR) \\
 &+ 13 \times h(IdxBond + 0.137116) * h(IR\_PC1\_EUR + 2.1965) * h(47.6094 - dVA) \\
 &+ ...
 \end{aligned}$$

avec  $h : x \mapsto \max(0, x)$

Les termes les plus influents dans le modèle MARS sont les mêmes que ceux ressortant dans le polynôme actuel, à savoir  $IdxBond$ ,  $Idx\_Eq$ ,  $dVA$  et  $IR\_PC1\_EUR$ .

---

1. Les coefficients ont été arrondis pour réduire la lourdeur d'écriture.

### 4.1.3 Principales différences entre les deux fonctions répliquatives

Tout d’abord, il est à noter que le modèle MARS utilise deux facteurs de risque qui ne sont pas pris en compte dans le polynôme actuel : Mort et MCAT, c’est à dire le risque de mortalité et le risque de mortalité subite dû à une catastrophe. Bien que l’importance de ces variables dans le modèle MARS reste faible, cela pourra avoir un impact en cas de stress important sur l’une et/ou l’autre de ces deux variables.

#### Comparaison de l’allure des proxies de manière univariée

Afin d’avoir une idée de l’allure que prennent les deux fonctions, dans le but de pouvoir les comparer visuellement, une analyse univariée a été effectuée afin d’obtenir des graphes en 2D représentant la forme des proxies sur un unique facteur de risque. Pour chacun des graphiques tous les facteurs de risque ont été mis à 0 excepté celui étudié qui varie sur l’intervalle d’apprentissage des modèles.

Ces graphes sont représentés figure 4.1, 4.2 et 4.3, la courbe bleue correspond au modèle MARS et la courbe rouge au polynôme actuel. En abscisse est représentée la variable étudiée et en ordonnée le *Stress On Liabilities* (SoL), stress sur le passif, correspondant à la différence entre le BE stressé et le BE central =  $SoL = BE - BE_0$ . Où  $BE_0$  est obtenu en mettant à 0 les stress sur tous les facteurs de risque.

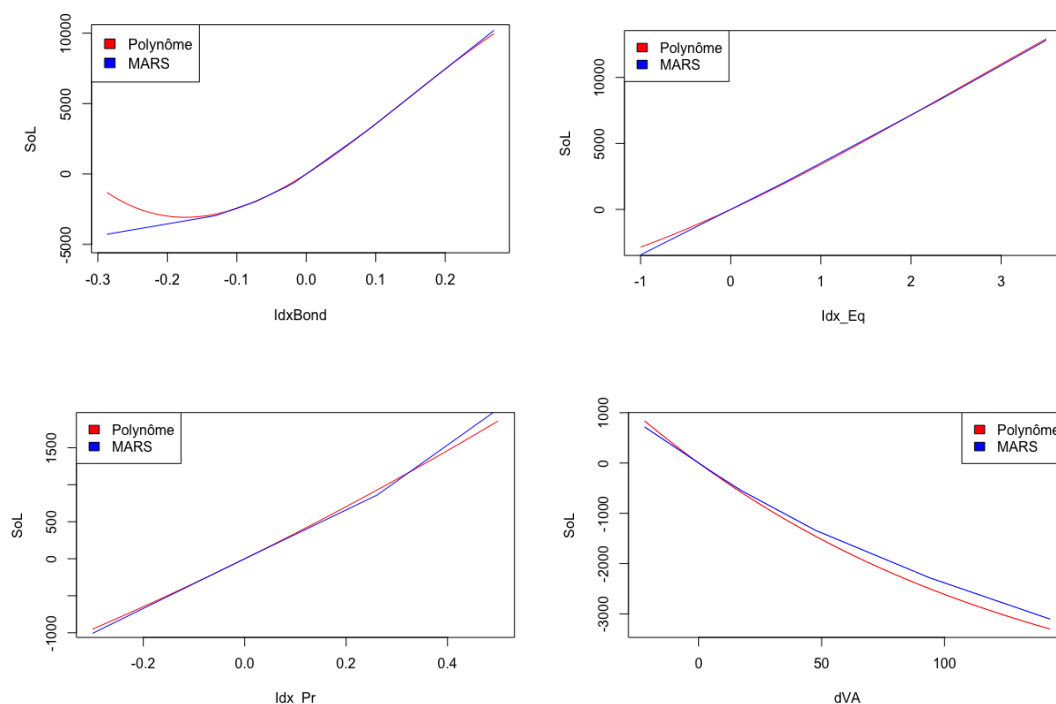


FIGURE 4.1 – Allure des proxies en 2D, pour chaque facteur de risque

Sur ces premiers graphiques, nous observons que les fonctions MARS et Polynôme sont très proches sur les variables  $Idx\_Eq$ ,  $Idx\_Pr$  et  $dVA$ . Quant à la variable  $IdxBond$ , les deux fonctions se confondent à partir de  $-0.1$  mais s'écartent considérablement pour des valeurs inférieures. En effet, la fonction MARS est croissante avec une pente qui s'accroît à partir de  $-0.1$ , puis encore plus à partir de  $0$ . Alors que la fonction polynomiale est décroissante jusqu'à  $-0.15$  puis devient croissante par la suite. Un stress sur  $IdxBond$  de  $0.3$  entraîne un écart d'environ 4 000 millions sur le *Stress on Liabilities*. De plus, en observant la limite droite du graphe, il est légèrement visible que la fonction polynomiale va avoir tendance à redevenir décroissante en dehors de l'intervalle d'apprentissage, alors que par sa définition la fonction MARS restera linéaire en dehors des bornes d'apprentissage.

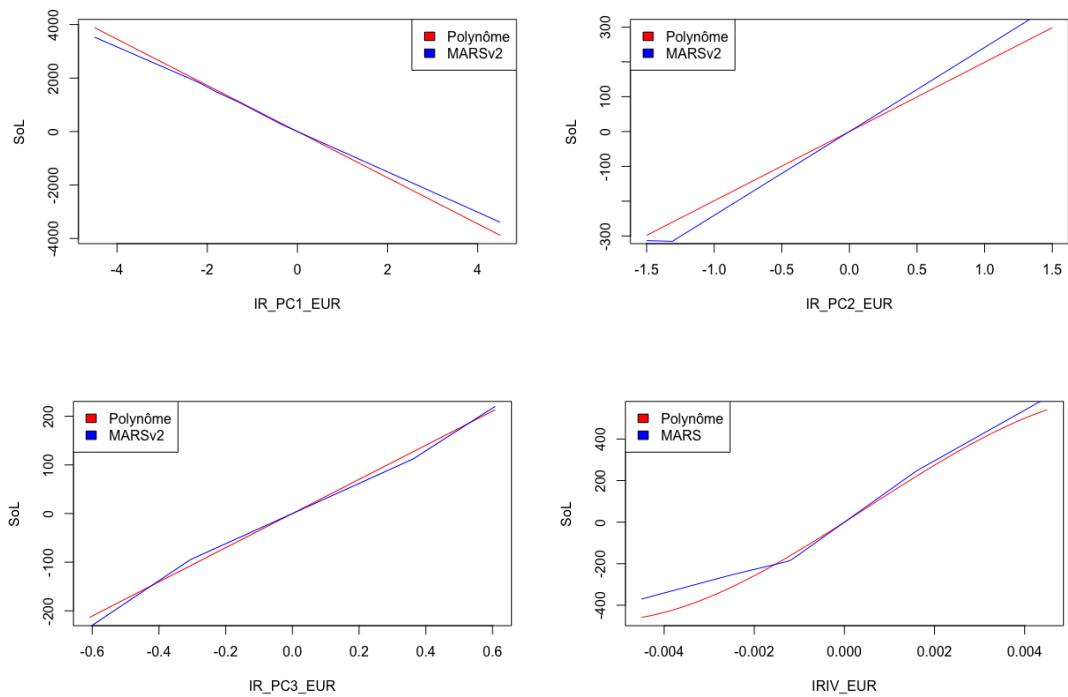


FIGURE 4.2 – Allure des proxies en 2D, pour chaque facteur de risque

Pour les variables  $IR\_PC1, PC2, PC3\_EUR$ , les deux fonctions sont globalement proches malgré la présence de certains noeuds sur la fonction MARS. De même il n'y a pas d'important écart sur la variable  $IRIV\_EUR$  mais il est notable que la fonction MARS est beaucoup moins lisse, dû au splines linéaires.

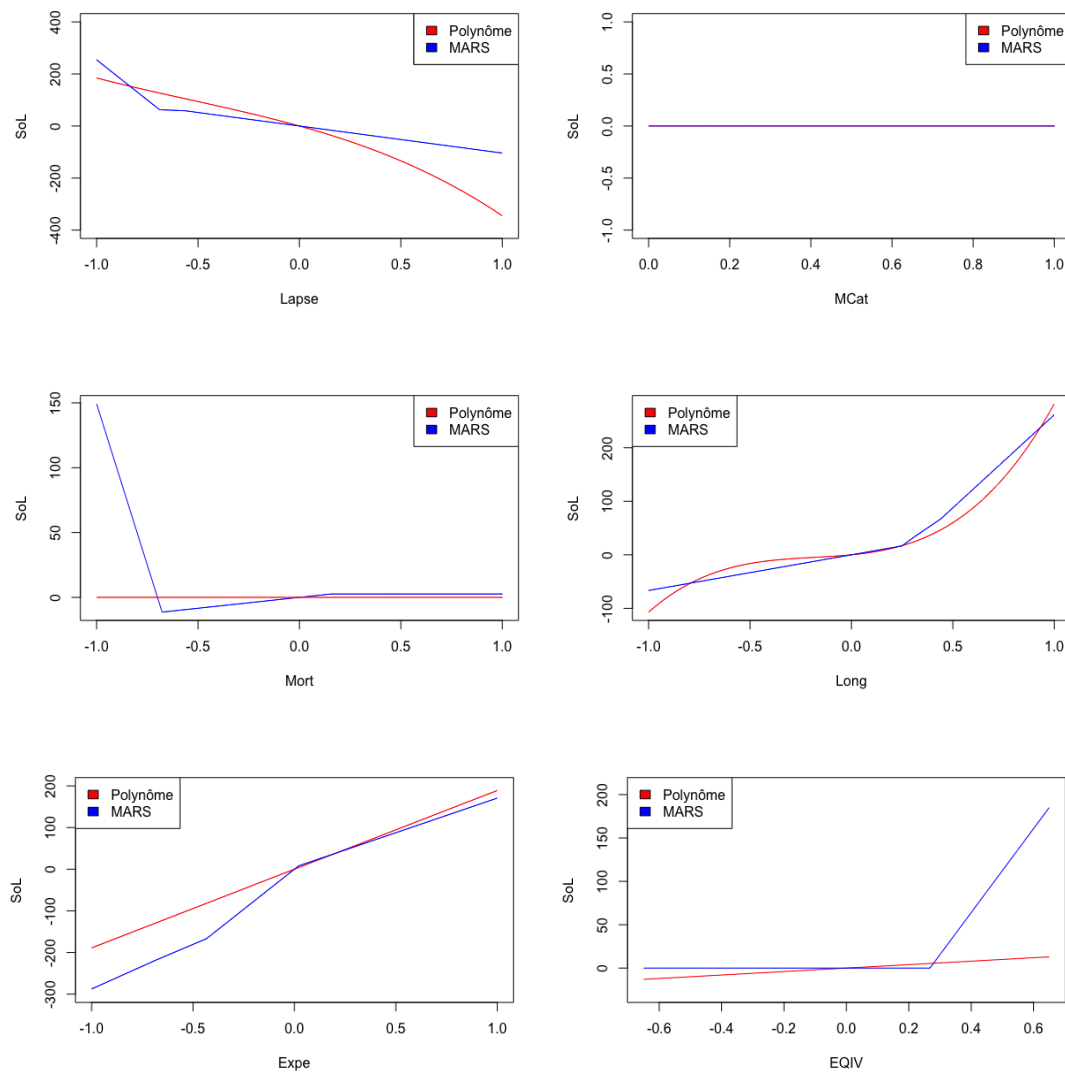


FIGURE 4.3 – Allure des proxies en 2D, pour chaque facteur de risque

Sur ces derniers graphes, de plus fortes divergences ressortent. Tout d'abord, il est intéressant de remarquer l'impact d'un choc négatif important sur la mortalité sur la fonction MARS, alors que ce facteur de risque n'est pas pris en compte dans le polynôme actuel. Ensuite, l'impact d'un stress positif élevé du facteur de risque EQIV est également beaucoup plus important sur la fonction MARS.

Du fait du processus de construction du modèle MARS par l'utilisation de noeuds et de splines linéaires, l'allure univariée de la fonction proxy ainsi créée présente de nombreuses cassures et est donc beaucoup moins lisse que la fonction paramétrique po-

ynomiale actuelle.

Bien que ces graphiques permettent une visualisation de l'allure des fonctions paramétriques des passifs stressés, il est important de garder en tête que de nombreux termes croisés sont présents dans les modèles et changent leur allure globale. En effet, en prenant l'exemple du graphe présentant l'impact univarié d'une variation du facteur de risque M<sub>cat</sub> il ressort que cette variable n'a aucun impact. Il s'avère que ce facteur de risque est bel et bien utilisé dans le modèle MARS mais seulement à l'intérieur d'un terme croisé s'annulant lorsque les autres variables sont nulles.

## 4.2 Estimation du SCR avec la forme paramétrique obtenue par la régression MARS

Dans cette section, nous allons calculer le SCR de la LoB contenant les produits d'assurance vie ayant une interaction ALM, constituée essentiellement d'épargne, que l'on obtiendrait en remplaçant la forme paramétrique polynomiale actuelle par le proxy du *Best Estimate* obtenu par le modèle MARS. Le SCR ainsi calculé sera ensuite comparé au SCR de cette ligne d'activité obtenu en utilisant le polynôme actuel, et les écarts observés seront analysés lors de la section suivante.

### 4.2.1 Processus de calcul

Afin de calculer le capital de solvabilité requis (SCR), comme il a été vu dans le Chapitre 2, nous utilisons la formule suivante :

$$SCR_0 = FP_0 - q_{0.5\%}(FP_{0+}) \quad (4.1)$$

avec  $FP_0$  le montant de Fonds Propres au temps  $t=0$ , c'est à dire les Fonds Propres initiaux,  $FP_{0+}$  les Fonds Propres au temps  $t=0^+$ .

Ou plus particulièrement, nous nous intéresserons à la distribution des pertes de Fonds Propres et le SCR sera obtenu comme suit :

$$SCR_0 = q_{99.5\%}(\Delta_{FP}) \quad (4.2)$$

avec  $\Delta_{FP} = FP_0 - FP_{0+}$  les pertes de Fonds Propres.

Or, au court de cette étude, ce ne sont pas les Fonds Propres en  $t=0^+$  mais les provisions *Best Estimate* qui ont été estimées par une fonction paramétrique. Il est alors possible d'accéder à la distribution des Fonds Propres par une soustraction des postes du bilan (*Balance Sheet Item*) de la manière suivante :

$$\Delta_{FP} = \Delta_{Actif} - \Delta_{BE} - \Delta_{GOBSI} + \Delta_{COBSI}$$

Où GOBSI et COBSI (Gross/Ceded Balance Sheet Item) correspondent aux autres postes du bilan, non détaillés dans le bilan prudentiel simplifié schématisé au Chapitre 1, et  $Actif_{0+}$  est valorisé par une approche de Curve Fitting.



Pour obtenir la distribution complète des Fonds Propres stressés il est alors nécessaire de connaître la distribution complète de tous les postes du bilan stressé en  $t=0^+$ . Pour cela,  $2^8$  (= 262 144) scénarios sont générés et rassemblés dans une table appelée *Risk Factor Scenario Set* (RFSS). Les fonctions d'approximation de chacun des postes du bilan sont alors appliquées à cet ensemble de  $2^8$  scénarios.

Seul le proxy du *Best Estimate* rentre dans le champs de cette étude mais, pour information, nous rappelons que la fonction paramétrique permettant d'estimer les actifs stressés est quant à elle obtenue par une approche Curve Fitting.

Afin de calculer le SCR de la ligne d'activité d'assurance vie que l'on obtiendrait si la fonction de régression du modèle MARS était utilisée dans le modèle interne pour estimer le *Best Estimate* stressé, il suffit d'appliquer la fonction paramétrique calibrée par le modèle MARS à l'ensemble des scénarios de la RFSS puis, ayant accès à la distribution complète du BE, il sera possible d'obtenir la distribution des Fonds Propres à l'aide des autres postes du bilan, et enfin nous serons capable de calculer le SCR correspondant par l'application d'un quantile.

$$SCR_0 = q_{99.5\%}(FP_0 - FP_{0+}) = q_{99.5\%}(FP_0 - Actif_{0+} - BE_{0+} - GOBSI_{0+} + COBSI_{0+})$$

Nous allons donc comparer les deux valeurs suivantes :

$$SCR_{MARS} = q_{99.5\%}(FP_0 - Actif_{0+} - f_{MARS}(RFSS) - GOBSI_{0+} + COBSI_{0+}) \quad (4.3)$$

et

$$SCR_{POLY} = q_{99.5\%}(FP_0 - Actif_{0+} - f_{POLY}(RFSS) - GOBSI_{0+} + COBSI_{0+}) \quad (4.4)$$

#### 4.2.2 Résultats obtenus

En effectuant la démarche présentée ci-dessus, en utilisant la fonction obtenue par l'algorithme MARS pour approcher le *Best Estimate* stressé de chacun des scénarios de la RFSS, nous obtenons un  $SCR_{MARS}$  de **1 941 millions d'euros**, sur la LoB regroupant les activités d'assurances vie présentant des interactions ALM.

### 4.3 Impact du changement de forme paramétrique

À travers cette section, le SCR de la LoB d'assurance vie calculé précédemment en utilisant la fonction calibrée par MARS pour valoriser le *Best Estimate* dans chacun des scénarios de la RFSS sera comparé au SCR de cette même ligne d'activité obtenu à partir du polynôme.

**NB :** Le SCR utilisé comme comparaison ne correspond pas tout à fait au SCR officiel de Generali pour cette ligne d'activité. En effet, lors du calcul du SCR, l'outil utilisé par

Generali intègre aux pertes de passifs ( $BE_0 - BE_{stress\ 0+}$ ) le risque de contrepartie et de défaut de crédit. De plus, un hair-cut (coefficient réducteur) sur l'effet positif du *Stochastic Volatility Adjustment* (SVA) est appliqué sur les pertes liées exclusivement à une variation de SVA. Cela vise, sur la demande de l'ACPR, à réduire l'effet positif du SVA en considérant les incertitudes et les approximations utilisées pour son calcul.

Par soucis de simplification, le SCR servant de comparaison sera alors calculé en suivant la méthode explicitée dans la section précédente et en utilisant la fonction polynomiale de cette LoB pour l'obtention de la distribution complète du *Best Estimate*, comme l'indique la formule (4.4).

### 4.3.1 Comparaison

#### SCR par la fonction polynomiale actuelle

Le capital de solvabilité requis pour la LoB unique, calculé en utilisant la fonction polynomiale actuelle pour obtenir la distribution complète du *Best Estimate* stressé en  $0^+$ ,  $SCR_{POLY}$  vaut **2 071 millions d'euros**, sur cette même ligne d'activité.

#### Écart entre les deux résultats de SCR

$SCR_{MARS}$	$SCR_{POLY}$	Écart	%
1 941 M	2 071 M	- 130 M	<b>-6.3%</b>

FIGURE 4.4 – Comparaison des SCR en fonction des fonctions répliquatives du BE utilisées

Comme le montre le tableau ci-dessus, le changement de fonction paramétrique pour la répliquative du *Best Estimate* stressé a un impact de 130 millions d'euros sur le calcul du SCR, soit une baisse de 6.3% du montant de capital de solvabilité requis en utilisant l'approche MARS plutôt que l'approche polynomiale.

Étant donné l'importance de l'impact, il est pertinent de s'intéresser à l'intégralité de la distribution des Fonds Propres pour chacune des méthodes.

Sur la figure 4.5 ci-dessous sont représentées les distributions empiriques des pertes de Fonds Propres après stress,  $FP_0 - FP_{stress\ en\ 0+}$ , calculées à partir du modèle MARS en bleu et à partir du polynôme en rouge. Sur ce graphique, nous remarquons que les queues de distribution des pertes de Fonds Propres calculées à partir du polynôme sont légèrement plus lourdes que celles des pertes de Fonds Propres calculées à partir de MARS. Cela pourrait être à l'origine de l'écart obtenu sur le SCR.

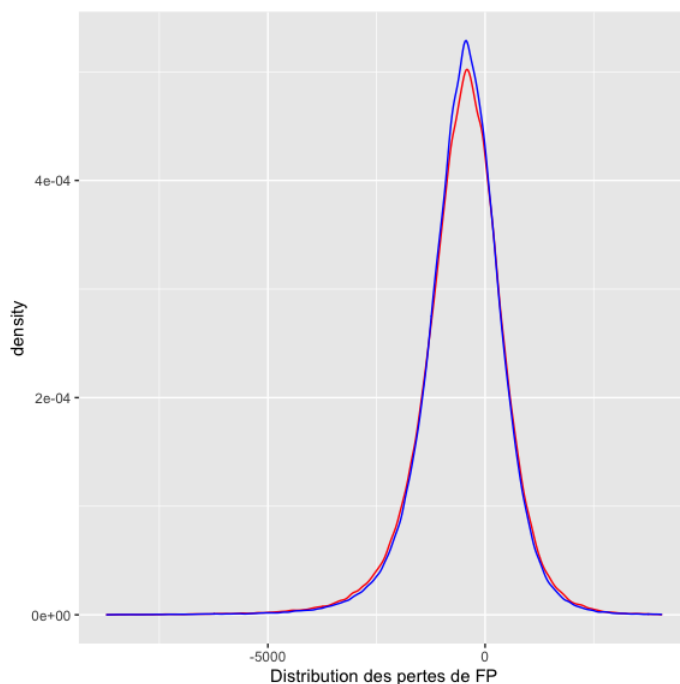


FIGURE 4.5 – Distribution des pertes de Fonds Propres stressés.  
En rouge est représentée la distribution par le polynôme et en bleu par MARS.

En termes de chiffres, le tableau suivant relate les quartiles ainsi que la moyenne et les extrémités des pertes de Fonds Propres obtenues par chacune des approches.

(En millions d'euros)	Pertes de $FP_{MARS}$	Pertes de $FP_{POLY}$
Minimum	-799 633	-1 158 320 314
1 <sup>er</sup> quartile	-986	-1 010
<b>Moyenne</b>	<b>-538</b>	<b>886 814</b>
Médiane	464	468
3 <sup>eme</sup> quartile	22	46
Maximum	87 268	231 411 042 996

FIGURE 4.6 – Étude de la distribution des pertes de Fonds Propres en fonction de l'approche utilisée

Tout d'abord, il ressort de ce tableau que les pertes de Fonds Propres calculées par l'approche polynomiale atteignent des extrêmes beaucoup plus importants. Cela a un impact direct sur la moyenne qui est alors très fortement supérieure à la moyenne des pertes de Fonds Propres par MARS, et également très éloignée de sa médiane. Quant à eux, les différents quartiles sont relativement proches avec les deux méthodes, bien que les valeurs obtenues par l'approche MARS aient l'air plus centrées, ce qui ressort sur le

graphique précédent représentant les deux fonctions de distribution des pertes de Fonds Propres.

L'existence de pertes de FP extrêmes avec la méthode polynomiale n'est pas forcément à l'origine de l'écart de SCR. En effet, étant donné que nous nous intéressons à un quantile, donc à un rang, quelques points aberrants peuvent n'avoir aucun impact sur la valeur du SCR; tant qu'ils n'atteignent pas la zone quantile (99.5% puisque nous nous focalisons directement sur les pertes de FP) et qu'ils ne modifient pas l'ordre des pertes de Fonds Propres. Il est néanmoins judicieux de s'intéresser à ces points extrêmes afin d'analyser s'ils peuvent être à l'origine d'un tel écart.

### 4.3.2 Analyse de l'écart observé

Une piste pouvant être à l'origine de cet écart de SCR en fonction de la fonction utilisée pour évaluer le *Best Estimate* stressé est la réaction des fonctions sur des scénarios dont les facteurs de risque se trouvent en dehors des zones d'apprentissage. Lorsque l'un des facteurs de risque d'un scénario de la RFSS (*Risk Factor Scenario Set*) sortira de l'intervalle sur lequel les fonctions paramétriques ont été calibrées, ce scénario sera appelé un scénario *Out of Range*.

#### Scénarios *Out of Range*

Dans la base d'apprentissage utilisée pour construire les fonctions de régression, chaque facteur de risque prend des valeurs comprises entre le quantile à 0.1% et le quantile à 99.9% de la distribution totale de ce facteur de risque. Les intervalles d'apprentissage de chacune des variables explicatives de la ligne d'activité d'assurance vie avec interactions ALM sont détaillés en Annexe C.

Étant donné que nous nous intéressons à un quantile, la valeur explosive de quelques termes extrêmes peut ne poser aucun problème, tant que ces scénarios sont exceptionnels et n'atteignent pas la zone du quantile à 99.5%. De plus, il faut que ces scénarios extrêmes ne changent pas l'ordre des valeurs de perte de Fonds Propres stressées pour que la valeur du SCR ne soit pas impactée.

Afin d'analyser dans quelle mesure l'explosivité de la fonction paramétrique polynomiale en dehors du champs d'apprentissage peut être responsable ou non de l'écart de SCR observé entre l'utilisation de MARS et du polynôme, une étude de ces scénarios *Out of Range* a été réalisée.

L'histogramme ci-dessous représente la proportion des scénarios *Out of Range* pour chaque facteur de risque, ainsi que le degré d'écart avec l'intervalle d'apprentissage. Plus un point sortira des rangs d'apprentissage de manière importante, et plus il sera représenté par une couleur foncée. Pour tenir compte de l'intensité d'écart avec l'intervalle d'entraînement, les points *Out of Range* ont été séparés en 4 niveaux. Notons

$$I_{app} = \prod_{i=1}^{Nb\ RF} [a_i, b_i]$$

avec  $a_i$  et  $b_i$  les bornes inférieure et supérieure de l'intervalle d'apprentissage du facteur de risque  $i$ . Nous définissons :

- **Niveau OoR 1** : Le choc sur un facteur de risque est en dehors de l'intervalle d'apprentissage mais ne dépasse pas le double des bornes de cet intervalle. Ce niveau est représenté par la couleur jaune.
- **Niveau OoR 2** : Le choc sur un facteur de risque  $i$  est en dehors de l'intervalle  $[2 \times a_i, 2 \times b_i]$  mais ne dépasse pas  $[5 \times a_i, 5 \times b_i]$ . Ce niveau est représenté par la couleur orange.
- **Niveau OoR 3** : Le choc sur un facteur de risque  $i$  est en dehors de l'intervalle  $[5 \times a_i, 5 \times b_i]$  mais ne dépasse pas  $[10 \times a_i, 10 \times b_i]$ . Ce niveau est représenté par la couleur rouge.
- **Niveau OoR 4** : Le choc sur un facteur de risque  $i$  est en dehors de l'intervalle  $[10 \times a_i, 10 \times b_i]$ . Ce niveau est représenté par la couleur noire.

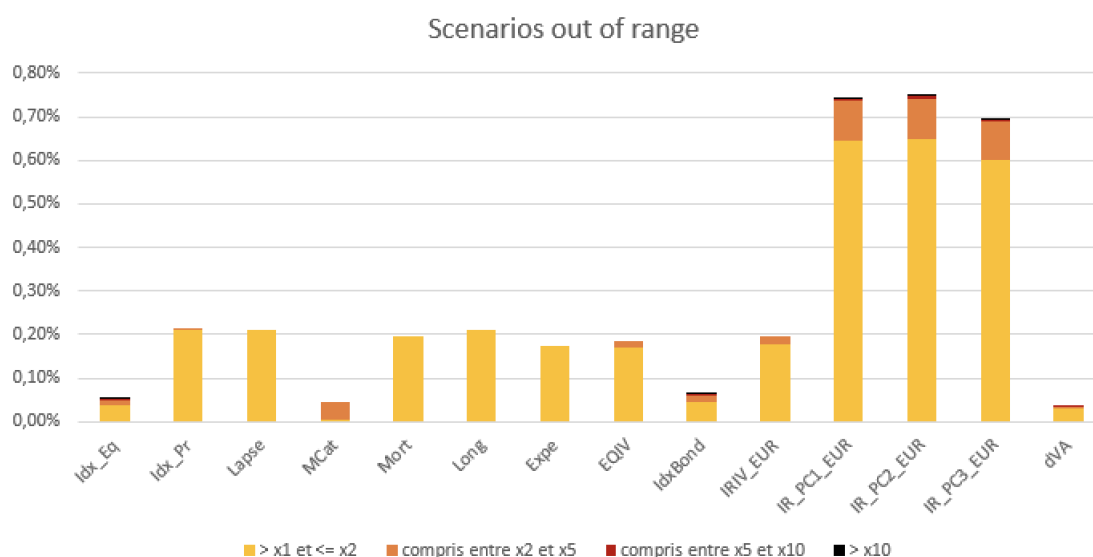


FIGURE 4.7 – Proportion des scénarios Out of Range sur les différentes variables explicatives

Sur cet histogramme, la première chose qui ressort est que le plus grand nombre d'*Out of Range* est observé sur variables de chocs sur le taux d'intérêt en euro : IR\_PC1, IR\_PC2 et IR\_PC3. En effet, les *Out of Range*, représentent en général moins de 0,2% des scénarios sur l'ensemble des facteurs de risque exceptés ces trois là, aux alentours de 0,7%. Les distributions des risques PC1, PC2 et PC3 ont été revues après la construction de la base de régression, ce qui explique qu'il y aient plus de 0.2% d'*Out of Range*.

Les chocs sur le risque de variation des marchés action et obligation (Idx\_Eq et IdxBond), font parti des facteurs de risque sortant le moins des rangs d'apprentissage. Il faut noter néanmoins que des *Out of Range* de niveau 4 sont observés sur ces variables, contrairement à la plupart des autres. Et que ce sont particulièrement sur les *Out of Range* de niveau 3 et 4 que des valeurs explosives sont observées.

Au total, il y a 9 430 scénarios comprenant au moins un facteur de risque *Out of Range*, soit 3.2 % de l'ensemble des scénarios. Parmi ces scénarios, 1 015 ( soit 0.39% de la RFSS) font partit d'un niveau 0oR supérieur ou égal à 2. La figure suivante permet de mettre en exergue la proportion des différents niveaux d'*Out of Range* parmi ces scénarios.

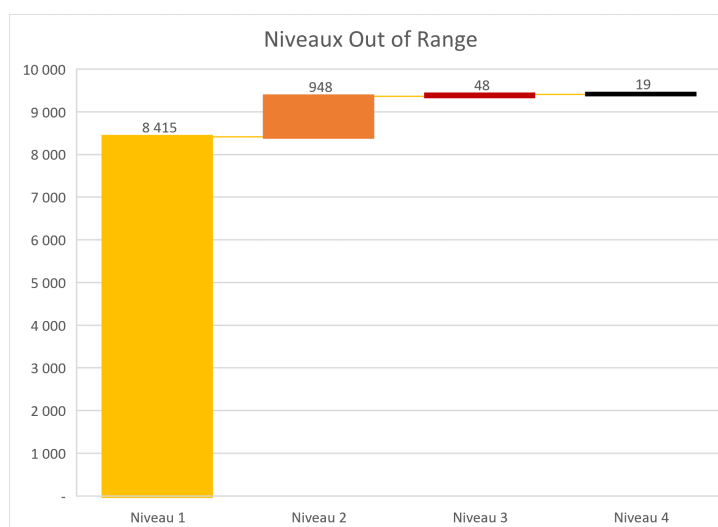


FIGURE 4.8 – Proportion des différents niveaux parmi les scénarios out of range.

Après analyse de ces scénarios, il est remarquable que les plus gros écarts de pertes de Fonds Propres observés correspondent à des scénarios *Out of Range* de niveau important, en particulier sur les variables IdxBond, IdxEq et les IR\_PC1/PC2/PC3. De plus, la définition d'un scénario *Out of Range* est assez sévère, ce qui amène à un taux de 3.2% très important (bien au delà de 0.05%). En réalité, lorsqu'un facteur de risque peu impactant sur l'estimation du *Best Estimate*, comme MCAT, Mort, Long, Expe et EQIV, sort de son intervalle d'apprentissage, cela n'a que très peu d'impact sur la valeur du *Stress on Liabilities* et ne conduit pas à des valeurs aberrantes.

Les scénarios étant *Out of Range* sur les variables IdxBond, IdxEq, Idx\_Pr, IR\_PC1\_EUR ou dVA sont du nombre de 2 745 soit 1,05% des données, et seulement 327 de ces scénarios sont d'un niveau *Out of Range* supérieur ou égale à 2, soit 0.12% de la RFSS.

### Réaction des fonctions paramétriques sur les scénarios Out of Range

Afin de mettre en exergue l'explosivité de la fonction paramétrique polynomiale lorsque les facteurs de risque sortent du périmètre d'entraînement, nous nous appuyons sur un des scénarios extrêmes de la RFSS. Ce scénario servant d'exemple prend les valeurs suivantes pour les facteurs de risque servant de variables explicatives aux modèles :

Eq	Pr	Lapse	MCat	Mort	Long	Expe	EQIV	Bond	IRIV	PC1	PC2	PC3	dVA
<b>12,7</b>	0,18	-0,06	0	-0,28	0,08	0,56	-0,22	<b>20,5</b>	-0,002	2,37	0,16	0,12	- 18

Les intervalles d'entraînement de chacun des facteurs de risque<sup>2</sup> sont présentés en Annexe C. Ce qui en ressort est que deux facteurs de risque sont *out of range* dans ce scénario : Idx\_Eq et IdxBond.

Pertes de FP avec MARS	Pertes de FP avec POLY	Ecart
$-3e^5$ M d'euros	$2.5e^9$ M d'euros	$2.49e^9$ M d'euros

FIGURE 4.9 – Pertes de Fonds Propres obtenues sur le scénario exemple en fonction de la méthode

Sur ce scénario, on observe un écart de  $2.5e^9$  M sur la valeur des pertes de Fonds Propres entre la méthode par MARS et la méthode polynomiale. Avec une perte de Fonds Propres égale à  $-3e^5$  M en utilisant la fonction obtenue par MARS, face à  $2,5e^9$  M en utilisant la fonction polynomiale. Outre l'écart important, nous constatons également que les deux résultats sont inversés en terme de placement dans la distribution, avec une perte de Fonds Propre très négative par MARS face à une perte extrêmement importante avec le polynôme.

Cet écart est facile à expliquer visuellement en s'intéressant à la réaction des fonctions sur les régions *Out of range* sur les deux facteurs de risque sortant de l'intervalle utilisé pour l'apprentissage : IdxBond et Idx\_Eq, qui sont les deux variables les plus impactantes sur la valeur de *Best Estimate* estimée, aussi bien avec MARS qu'avec le polynôme.

2. C'est à dire les intervalles dans lesquels les facteurs de risque des scénarios de la base d'apprentissage étaient représentés.

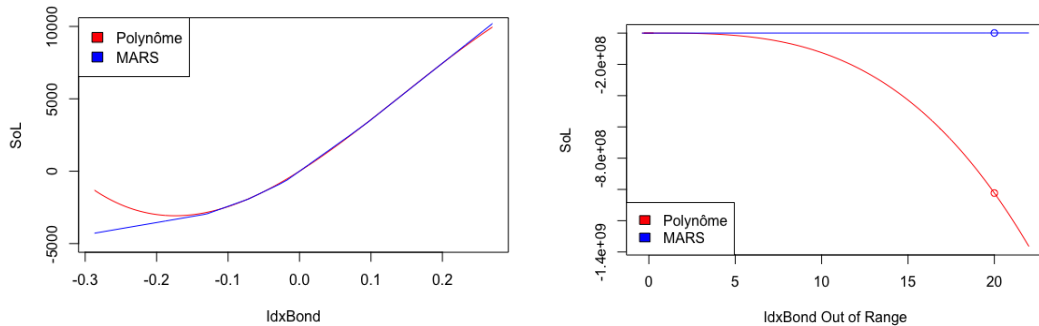


FIGURE 4.10 – Réaction des proxies en dehors des régions d'apprentissage de IdxBond

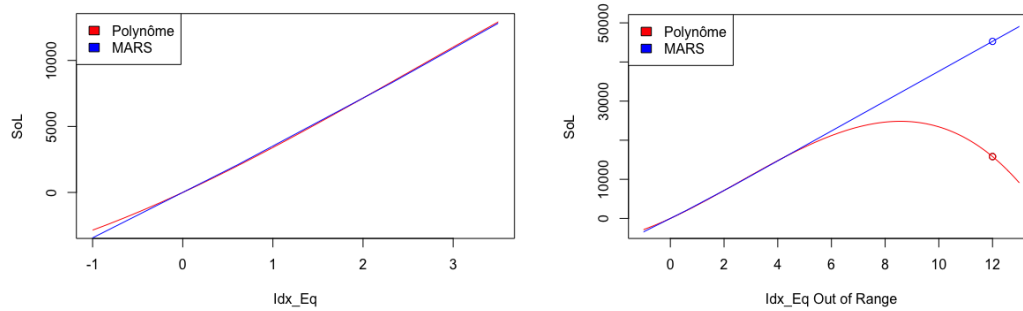
Dans la base d'apprentissage, à partir de laquelle les modèles ont été calibrés, le stress sur le facteur de risque IdxBond était compris entre -0,28 et 0,26. Or dans la RFSS (*Risk Factor Scenario Set*), les stress sur ce facteur de risque peuvent sortir énormément de cette région d'apprentissage, comme c'est le cas sur ce scénario pris en exemple pour lequel le stress sur IdxBond est égal à 20,5.

Le modèle MARS a la particularité de rester linéaire en dehors des zones d'apprentissage, ce qui permet une certaine cohérence avec la tendance observée et ce qui empêche d'atteindre des impacts trop explosifs. Quand à lui, le polynôme peut exploser comme c'est le cas ici, dû à un terme en  $-IdxBond^3$ . La figure 4.10 permet de mettre en évidence cela avec à gauche un graphique représentant la fonction polynomiale (en rouge) et par MARS (en bleu) pour des stress sur IdxBond compris dans l'intervalle d'apprentissage, et à droite l'allure de ces fonctions en dehors du périmètre d'entraînement. Les points correspondent à la valeur du *Stress on Liabilities* ( $BE_0 - BE_{stress0+}$ ) pour un IdxBond égale à 20,5.

L'allure du polynôme en dehors des rangs d'apprentissage est assez incohérente. En effet, dans le champs d'entraînement les deux modèles captaient une tendance à la hausse du *Stress on Liabilities* lorsque le choc sur IdxBond augmentait. Or, lorsque le choc sur IdxBond continue d'augmenter et sort de l'intervalle d'apprentissage, la fonction polynomiale prédit un *Stress on Liabilities* décroissant en fonction du choc sur IdxBond. Le comportement de la fonction MARS, restant linéaire selon la même droite que la dernière portion du périmètre d'apprentissage, paraît plus pertinent.

Ce scénario particulier expose un réel problème de la fonction polynomiale étant donné que la valeur de *Stress on Liabilities* prédite n'est pas seulement extrême mais également de sens inverse à ce à quoi nous nous attendrions au vu de ce qui est observé sur l'intervalle d'apprentissage. Ce changement de place dans la distribution a un impact sur le quantile à 99.5%, puisqu'il impliquera un décalage du scénario quantile.



FIGURE 4.11 – Réaction des proxies en dehors des régions d'apprentissage de  $Idx\_Eq$ 

Le stress sur  $Idx\_Eq$  sort lui aussi totalement de l'intervalle d'apprentissage avec une valeur de 12.7, alors que les points ayant servi à construire les fonctions paramétriques étaient compris entre -0.77 et 3.45. Sur la figure ci-dessus, nous observons encore que plus la valeur du stress s'éloigne du champ d'apprentissage et plus les valeurs du polynôme et de la fonction obtenue par MARS s'écartent.

Une fois de plus, la valeur du SoL (*Stress on Liabilities*) diminue à partir d'un certain moment dû à un terme négatif avec  $Idx\_Eq$  au cube lors de l'utilisation du polynôme. Le modèle MARS reste linéaire en dehors des rangs d'apprentissage, et est plus cohérent avec la tendance à la hausse observée sur le périmètre d'entraînement.

CONCLUSION DU CHAPITRE :

Les fonctions paramétriques obtenues par MARS ou par la régression polynomiale actuelle sont globalement proches sur l'intervalle d'apprentissage sur lequel elles ont été construites, mais diffèrent sur plusieurs points. Tout d'abord MARS utilise deux facteurs de risque non pris en compte dans le polynôme : Mort et M<sub>Cat</sub>. De plus, par construction de l'algorithme MARS, l'allure de la fonction de régression obtenue est linéaire par morceaux avec plusieurs points de coupure au niveau des noeuds, le polynôme aura quant à lui une forme beaucoup plus lisse.

Le SCR de la ligne d'activité d'assurance vie comprenant des interactions ALM a été calculé en utilisant chacune des fonctions paramétriques pour l'obtention de la distribution complète des passifs stressés. Un impact de - 6.3%, soit - 130 millions d'euros est calculé lorsque la fonction de MARS est utilisée. Afin d'essayer d'expliquer cet écart significatif, les scénarios *Out of Range* ont été étudiés.

En effet, lorsque les valeurs des chocs sortent du périmètre sur lequel les deux fonctions ont été construites, les écarts entre elles s'accroissent. L'allure du polynôme en dehors des intervalles d'apprentissage est incontrôlable et peut amener à des résultats aberrants. La fonction de MARS restera quant à elle linéaire une fois sortie de l'intervalle d'entraînement, ce qui permet de garder une certaine cohérence avec la tendance observée sur les données d'apprentissage, et de ne pas atteindre des valeurs trop explosives.

# Chapitre 5

## Conclusion

### 5.1 Conclusion

Dans ce mémoire, nous nous sommes intéressés à la projection de la distribution complète des engagements *Best Estimate* stressés, dans une optique de calcul de capital de solvabilité requis par modèle interne.

Compte tenu des interactions ALM présentes dans les contrats d'assurance vie, les engagements *Best Estimate* doivent être valorisés de manière stochastique. De ce fait, l'obtention de la distribution complète du *Best Estimate* à 1 an est complexe, puisqu'elle requiert la réalisation de calculs stochastiques imbriqués.

Il a été vu que la méthode de calcul la plus naturelle permettant d'obtenir la distribution complète des passifs à 1 an, est la méthode des Simulations dans les Simulations. Mais, dû aux problèmes de temps de calculs inhérents à cette méthode, différentes approches alternatives ont été abordées : l'accélérateur de SdS, le *Replicating Portfolio* ou encore des approches paramétriques comme le *Curve Fitting* et le *Least Square Monte Carlo* (LSMC).

Nous nous sommes particulièrement intéressés à la méthode *Least Square Monte Carlo*. Le principe de cette méthode est de construire une fonction paramétrique du *Best Estimate* stressé en fonction des facteurs de risque identifiés, afin de valoriser les engagements *Best Estimate* directement, évitant ainsi de nombreuses simulations dans le modèle de projection ALM ce qui induit un gain important en terme de temps de calcul. Pour construire cette fonction paramétrique, l'approche LSMC consiste à calculer des valeurs inexactes de *Best Estimate* sur de nombreuses simulations primaires, par la réalisation de seulement 2 simulation Monte Carlo secondaires, puis d'effectuer une régression de ces valeurs inexactes.

Actuellement, la régression appliquée à ces valeurs est une régression polynomiale, dont les termes ont été rigoureusement sélectionnés par un algorithme *Adaptive Stepwise Backward* de sélection de termes. L'objectif de cette étude a été de mettre à l'épreuve cette régression par des modèles de *Machine Learning*. Pour cela, quatre modèles ont été calibrés sur une base de données issue de l'approche LSMC (contenant pour 25 000 scénarios).

rios primaires des valeurs approchées de BE obtenues par la réalisation de 2 simulations antithétiques de Monte Carlo ) :

- Forêt aléatoire,
- *Extreme Gradient Boosting*,
- *Multivariate Adaptive Regression Splines* (MARS),
- *Support Vector Regression*.

Après avoir effectué une comparaison des performances de ces modèles sur une base de validation contenant 125 valeurs exactes de *Best Estimate* (calculées par 1 000 simulations Monte Carlo), nous avons constaté que MARS surpasse les trois autres modèles. De plus, MARS améliore légèrement les performances de la régression sur la ligne d'activité regroupant les contrats d'assurance vie avec interactions ALM comparativement au polynôme actuel.

Nous nous sommes ensuite intéressés à l'impact qu'aurait un changement de forme paramétrique du *Best Estimate* stressé dans le calcul du SCR de la compagnie. Pour cela, nous avons appliqué la fonction de régression de MARS à  $2^{18}$  scénarios économiques afin d'obtenir la distribution complète du BE stressé. À partir de cette distribution, ainsi que des distributions des autres postes du bilan prudentiel, nous avons pu déduire la distribution complète des Fonds Propres stressés de l'entreprise, et ainsi son SCR en calculant le quantile à 0.5% de cette distribution estimée.

Bien que les régressions par MARS et polynomiale soient très proches sur les intervalles d'apprentissage, avec notamment des performances similaires sur la base de validation, l'utilisation de l'approche MARS pour valoriser le BE stressé plutôt que le polynôme conduit à une diminution de 130 millions d'euros (-6%) du SCR.

Une source d'explication de cet écart considérable est la réaction des deux fonctions paramétriques en dehors des intervalles d'apprentissage. En effet, les fonctions de régression ayant été construites à partir de scénarios économiques dont chacun des facteurs de risque était compris entre le quantile 0.1% et 99.9% de leur distribution, certains des scénarios utilisés pour construire la distribution complète du BE stressé sortent des rangs d'apprentissage. En dehors de ces intervalles d'apprentissage, les deux fonctions ont des réactions totalement différentes. Le polynôme est incontrôlable et peut atteindre des valeurs extrêmes très importantes dû à la présence de puissances élevées de certains facteurs de risque. Quant à elle, la fonction de régression obtenue par MARS a l'avantage de rester linéaire en dehors des zones d'apprentissage. Cela permet de rester en cohérence avec les tendances observées sur les données d'apprentissage, et de ne pas amener à des incohérences.

L'existence de points extrêmes incohérents obtenus par le polynôme sur certains scénarios Out of Range était connu, mais il était supposé que leur nombre n'atteignait pas la zone quantile. Cette hypothèse s'est révélée correcte puisque ces scénarios extrêmes

représentent seulement une quinzaine de scénarios. Néanmoins, le comportement parfois incohérent du polynôme en dehors des zones d'apprentissage implique parfois un changement de l'ordre des pertes de fonds propres, ce qui impacte la valeur du SCR.

Cette étude nous a permis de mettre en exergue une grosse limite de la régression polynomiale utilisée actuellement au sein de la méthode LSMC pour le calibrage d'une fonction paramétrique du BE stressé, qui est son comportement en dehors des zones d'apprentissages.

L'approche par MARS ne sera pas utilisée pour le calcul officiel du SCR, mais pourra servir de comparaison. Malgré le gain important en terme de SCR qu'elle pourrait apporter, les performances de la régression par MARS ne surpassent pas celles des polynômes sur toutes les lignes d'activités, en particulier sur la retraite. De plus, l'approche par régression polynomiale a le mérite d'être plus prudente.

## 5.2 Pour aller plus loin

Plusieurs idées permettant d'approfondir cette étude seront abordées dans cette section.

### Utilisation de splines cubiques

Malgré les bons résultats obtenus par l'algorithme MARS, qui pour rappel utilise des splines linéaires et des termes d'interactions entre les différentes variables explicatives, il pourrait être intéressant de modifier l'algorithme MARS de telle sorte qu'il n'utilise plus des splines linéaires, mais des splines cubiques. En plus de la possibilité d'améliorer les performances du modèle, cela permettrait d'obtenir une fonction paramétrique plus lisse, ne présentant pas les cassures observées sur l'allure de MARS.

Néanmoins, le recours à des splines cubiques pourrait amener le modèle à présenter les mêmes limites que celles observées sur le polynôme. De plus, remplacer les splines linéaires par des splines cubique augmenterait le temps de fitting de la régression.

### Choix de la base d'apprentissage des modèles

Dans cette étude nous avons remis en cause l'étape régression au sein de la méthode Least Square Monte Carlo par l'utilisation de machine learning, pour cela la base de données permettant aux modèles de s'entraîner était construite par une approche LSMC. Ce choix a été essentiellement opérationnel, Generali utilisant cette approche dans son modèle interne pour créer les fonctions paramétriques polynomiales répliquant le *Best Estimate* stressé, nous avons accès directement à cette base de données. Mais pourquoi ne pas challenger la méthode LSMC elle-même ?

En effet le choix de cette base de données d'apprentissage pourrait être remis en question afin d'utiliser d'autres modèles de régression, pouvant être plus performants sur

une base contenant un nombre plus restreint de points, mais qui seraient plus précis, comme c'est le cas avec l'approche de Curve Fitting.

Par exemple, dans son mémoire, HOUKONNOU Prudence [14] utilise un modèle de krigeage stochastique[16] pour améliorer la régression mise en place au sein d'une approche LSMC permettant d'estimer le *Best Estimate* stressé. L'intérêt de cette méthode est qu'elle permet de prendre en compte l'inexactitude des points de la base d'apprentissage due au faible nombre de simulations Monte Carlo effectuées, en tenant compte de la variance observée entre les sommes actualisées des cash-flows futurs obtenues à chaque simulation secondaire, pour un même scénario primaire.

La limite de cette méthode, qui est la cause pour laquelle elle ne fut pas utilisée dans notre étude, est que sa complexité est en fonction du nombre de points de la base de données d'apprentissage. L'utilisation de cette méthode sur notre base d'apprentissage contenant 25 000 points était irréalisable opérationnellement<sup>1</sup>.

Néanmoins, il fût très intéressant de remarquer que le modèle de krigeage stochastique construit à partir de seulement 10% des données tirées aléatoirement, soit 2 500 points de *Best Estimate* approchés, nécessitant 11min de calcul, obtenait de très bon résultats (au regard du peu de données utilisées) avec notamment une *Weighted Relative Error* (WRE) égale à 0.42%<sup>2</sup>.

Puisque la limite opérationnelle d'un modèle de krigeage stochastique est son temps de calcul, explosant lors de l'utilisation d'un nombre important de points d'apprentissage, mais que le modèle obtient de bon résultats à partir de peu de données, une idée serait de tester les performances de ce modèle lors de l'étape de régression d'une approche Curve Fitting. En effet, l'approche Curve Fitting va réduire le nombre de scénarios primaires, puis effectuer une valorisation Monte Carlo du *Best Estimate* avec un nombre conséquent de simulations risque neutre, la base de donnée utilisée pour la régression est donc considérablement plus petite.

Dans le même esprit, d'autres modèles pourraient être testés en partant d'une base d'apprentissage différente de celle utilisée dans cette étude.

---

1. 1.5 jours de calcul ont été nécessaires pour construire un modèle de krigeage stochastique sur la moitié des données, soit 12 500 points. De plus, la complexité n'étant pas linéaire, cela demanderait de nombreux jours pour calibrer un modèle sur l'ensemble des données

2. Résultat obtenu en sélectionnant aléatoirement 10% des données comme base d'apprentissage, avec l'utilisation de la fonction de covariance Mattern 3\_2 pour les processus gaussiens simulés.

## Annexe A

# Fonctionnement d'un arbre de décision pour la régression

### Principe général

Le principe d'un arbre de décision est de séparer les individus d'une base de données future en différentes feuilles (groupes de données final) distinctes les plus homogènes possible, en apprenant sur des données dont la variable cible est connue.

Un arbre commence par une racine (tous les individus) puis se sépare en différentes branches afin de finir par des feuilles. On appelle *noeuds* les différents embranchements de l'arbre. A chaque noeud a lieu un test sur une variable explicative, qui va permettre de séparer les individus en plusieurs groupes en fonction de la réponse du test. Chacune des réponses possibles du test donnera une branche.

Par exemple, le premier test pourrait être : « $\text{IdxBond} > 0.2?$ ». Alors deux branches apparaîtront et sépareront nos données en 2 groupes avec dans chacun d'entre eux une certaine variance de la variable à expliquer («BEGRO»).

Ainsi de suite, aux noeuds suivants, d'autres tests sur d'autres variables explicatives sont effectués dans le but d'obtenir des feuilles les plus homogènes possibles, c'est à dire dont la variance est minimale.

L'arbre est construit de sorte à ce que les tests effectués soient ceux permettant de séparer un maximum les individus de la bases de départ en différente groupes. En pratique, le choix de la segmentation est effectué par l'optimisation d'une métrique.

Plus une variable est capable de séparer un grand nombre d'individus en différents groupes de variances inférieure, plus cette variable est importante.

Une fois l'arbre créé, pour l'appliquer à d'autres données il suffit de parcourir les différentes branches, en partant de la racine (première segmentation des données) jusqu'aux feuilles. La feuille dans laquelle termine l'individu déterminera la valeur de la variable à prédire : ce sera la moyenne des variables cibles des différents individus d'apprentissage présents dans cette feuille.

## Annexe B

# Intervalles de *fitting* pour les modèles calibrés sur la ligne d'activité assurance vie

Les intervalles de fitting sont déterminés par Generali de telle sorte à aller jusqu'au quantile 0.01% et 99.9% de la distribution de chaque variable. Le tableaux suivant résume les bornes inférieures et supérieures de chaque variable utilisée pour le *fitting* du polynôme de la ligne d'activité d'assurance vie.

<b>Variables</b>	<b>Borne Inf</b>	<b>Borne Sup</b>
IdxEq	-0.78	3.46
IdxPr	-0.31	0.44
Lapse	-1	1
MCat	0	1
Mort	-1	1
Long	-1	1
Expe	-1	1
EQIV	-0.64	0.64
IdxBond	-0.29	0.27
IRIV_EUR	-0.00405	0.00405
IR_PC1_EUR	-4.27	4.27
IR_PC2_EUR	-1.47	1.47
IR_PC3_EUR	-0.61	0.61
dVA	-22	143



## Annexe C

# Construction de la *Risk Factor Scenario Set* (RFSS)

Le processus de génération de la *Risk Factor Scenario Set* (RFSS) est divisé en trois phases :

1. Un grand nombre (actuellement  $2^{18}$ ) de nombres aléatoires non corrélés est produit pour chaque *risk-driver* en utilisant le générateur de nombres aléatoires choisi (actuellement des suites SOBOLS). Cela produit une matrice de taille  $n \times d$ , avec  $n$  le nombre de scénarios et  $d$  le nombre de *risk-driver*, contenant des nombres non corrélés quasi-aléatoires selon une loi uniforme sur  $[0,1]$ .
2. Les  $d$  variables aléatoires pour chaque scénario sont ensuite transformées en utilisant la copule<sup>1</sup> adéquate (copule gaussienne actuellement). La transformation requiert l'application de la décomposition de Cholesky<sup>2</sup> à la matrice de corrélation souhaitée, c'est pour cela qu'une matrice symétrique définie positive est nécessaire à ce stade. La copule impose la structure de dépendance souhaitée sur les échantillons, ce qui entraîne la production du *risk-drivers scenarios set*(RDSS), dans lequel tous les *risk-drivers* suivent une distribution normale  $(0,1)$ , et reflète la structure de dépendance entre les risques.
3. La *Risk-factors scénario set*(RFSS) est alors créée en appliquant la distribution marginale pour chaque *Risk-factor* au *Risk-driver* approprié, pour produire une RFSS dans laquelle tous les *Risk-factors* ont la distribution marginale choisie. Pour

---

1. Une copule est fonction du cube unité  $C : [0, 1]^d \rightarrow [0, 1]$  qui satisfait les propriétés suivantes :

- Soit  $x \in [0, 1]^d$ ,  $C(x) = 0$  si il existe  $i \in [1, d]$  tel que  $x_i = 0$ .
- Soit  $x \in [0, 1]^d$ ,  $C(x) = 1$  si  $\forall i \in [1, d]$   $x_i = 1$ .
- $C$  augmente de façon monotone  $\forall x_i \in [0, 1]^d$ .

2. **Théorème ( décomposition de Cholesky )** : Soit  $A$  une matrice symétrique définie positive. Alors il existe une matrice triangulaire inférieure  $L$  telle que  $A = L^t L$ . De plus, si l'on impose aux coefficients diagonaux de  $L$  d'être positifs, cette factorisation est unique.

cela, la distribution normale[0,1] des *Risk-drivers* est inversée puis l'inverse de la fonction de répartition marginale  $F^{-1}(x)$  de la distribution marginale choisie pour chaque *Risk-factor* est appliquée au *Risk-driver*. Dans cette étape, il est possible d'associer plusieurs *Risk-factors* à un seul *Risk-drivers*, ce qui donne une RFSS plus grande que la RDSS.

## Annexe D

# Résultats du modèle MARS sur les autres lignes d'activités

L'intégralité du rapport de ce mémoire s'est concentré sur la ligne d'activité contenant les contrats d'assurance vie présentant des interactions actif/passif. La comparaison des résultats obtenus par la méthode MARS face aux résultats des polynômes actuels a également été faite sur d'autres lignes d'activités. Ces résultats seront présentés ci-dessous.

### D.1 Ligne d'activité regroupant les contrats de protection : LoB PRO

Afin d'expliquer au mieux la provision *Best Estimate* stressée de la ligne d'activité regroupant les contrats de protection, les facteurs de risques suivants ont été considérés comme variables explicatives au modèle :

1. Morb
2. HCl<sub>a</sub>
3. IR\_PC1\_EUR
4. dVA
5. Mort
6. M<sub>Cat</sub>
7. IR\_PC2\_EUR
8. IR\_PC3\_EUR
9. Expe

10. Lapse

11. Long

Les hyper-paramètres de la fonction *earth()* du logiciel R ont été déterminés par validation croisée et sont les suivants :

- **degree** = 5 ;
- **nk** = 250 ;
- **endspan** = 1400 ;
- **fast.k** = 20.

Le tableau figurant ci-dessous présente les performances du modèle MARS comparativement au polynôme actuel sur la LoB Protection. Ces résultats ont été obtenus sur une base de validation contenant 125 valeurs de *Best Estimate* précise, calculée à partir de 1 000 simulations risque neutre chacune.

LoB PRO	Polynôme ac- tuel	MARS
RMSE	53	<b>50</b>
WRE	0.73%	0.75%
Erreur relative abso- lue max	2.76%	<b>2.13%</b>
R <sup>2</sup>	0.969	<b>0.973</b>

FIGURE D.1 – Résultats obtenus sur la base de validation de la LoB Protection

## D.2 Ligne d'activité regroupant les contrats de retraite : LoB PENUN

Afin d'expliquer au mieux la provision *Best Estimate* stressée de la ligne d'activité regroupant les contrats de retraite (PERP), les facteurs de risques suivants ont été considérés comme variables explicatives au modèle :

1. IR\_PC1\_EUR
2. Idx\_Eq
3. IdxBond
4. dVA
5. Long

6. IR\_PC2\_EUR
7. EQIV
8. IR\_PC3\_EUR
9. Idx\_Pr
10. Expe
11. IRIV
12. Lapse

Les hyper-paramètres de la fonction *earth()* du logiciel R ont été déterminés par validation croisée et sont les suivants :

- **degree** = 4;
- **nk** = 150;
- **endspan** = 500;
- **fast.k** = 19.

Le tableau figurant ci-dessous présente les performances du modèle MARS comparativement au polynôme actuel sur la LoB PENUN. Ces résultats ont été obtenus sur une base de validation contenant 125 valeurs de *Best Estimate* précise, calculée à partir de 1 000 simulations risque neutre chacune.

LoB PENUN	Polynôme ac- tuel	MARS
RMSE	<b>80.3</b>	149
WRE	<b>0.32%</b>	0.60%
Erreur relative abso- lue max	<b>1.07%</b>	1.85%
R <sup>2</sup>	<b>0.9973</b>	0.9905

FIGURE D.2 – Résultats obtenus sur la base de validation de la LoB *pension*

### D.3 Ligne d'activité regroupant les contrats de retraite professionnelle : LoB PENTR

Afin d'expliquer au mieux la provision *Best Estimate* stressée de la ligne d'activité regroupant les contrats de retraite professionnel (RFFI), les facteurs de risques suivants ont été considérés comme variables explicatives au modèle :

1. Idx\_Eq
2. IdxBond
3. Long
4. IR\_PC1\_EUR
5. dVA
6. Expe
7. Idx\_Pr
8. IR\_PC2\_EUR
9. IRIV\_EUR
10. IR\_PC3\_EUR
11. EQIV
12. Lapse
13. HCl<sub>a</sub>

Les hyper-paramètres de la fonction *earth()* du logiciel R ont été déterminés par validation croisée et sont les suivants :

- **degree** = 3;
- **nk** = 100;
- **endspan** = 1200;
- **fast.k** = 18.

Cela amène à construire un modèle MARS contenant 55 termes.

Le tableau figurant ci-dessous présente les performances du modèle MARS comparativement au polynôme actuel sur la LoB PENTR. Ces résultats ont été obtenus sur une base de validation contenant 125 valeurs de *Best Estimate* précise, calculée à partir de 1 000 simulations risque neutre chacune.

LoB PENTR	Polynôme actuel	MARS
<b>RMSE</b>	<b>1.57</b>	2.02
<b>WRE</b>	<b>0.14%</b>	0.17%
<b>Erreur relative absolue max</b>	<b>0.47%</b>	0.59%
<b>R<sup>2</sup></b>	<b>0.9997</b>	0.9995

FIGURE D.3 – Résultats obtenus sur la base de validation de la LoB *pension*

## D.4 Ligne d'activité regroupant les contrats en unités de compte : LoB UC

Afin d'expliquer au mieux la provision *Best Estimate* stressée de la ligne d'activité regroupant les contrats en unités de compte, les facteurs de risques suivants ont été considérés comme variables explicatives au modèle :

1. Idx\_Eq
2. IdxBond
3. Idx\_Pr
4. Lapse
5. Expe
6. IR\_PC1\_EUR
7. Mort
8. dVA
9. EQIV
10. IR\_PC2\_EUR
11. Mort
12. IRIV\_EUR
13. IR\_PC3\_EUR

Les hyper-paramètres de la fonction *earth()* du logiciel R ont été déterminés par validation croisée et sont les suivants :

- **degree** = 3;
- **nk** = 150;
- **endspan** = 1100;
- **fast.k** = 20.

Le tableau figurant ci-dessous présente les performances du modèle MARS comparativement au polynôme actuel sur la LoB UC. Ces résultats ont été obtenus sur une base de validation contenant 125 valeurs de *Best Estimate* précise, calculée à partir de 1 000 simulations risque neutre chacune.

LoB UC	Polynôme ac- tuel	MARS
RMSE	8	9.5
WRE	0.028%	0.033%
Erreur relative abso- lue max	0.097%	<b>0.082%</b>
R <sup>2</sup>	0.999999	0.999998

FIGURE D.4 – Résultats obtenus sur la base de validation de la LoB des unités de compte



# Glossaire

**AIC** Critère d'information d'Akaike.

**BE** Provisions Best Estimate.

**Fitting Time** Temps nécessaire de calibration du modèle, une fois les hyper-paramètres renseignés.

**FP** Fonds Propres.

**LoB** Line of Business, Ligne d'activité.

**LSMC** Least Square Monte Carlo.

**MARS** Multivariate Adaptive Regression Spline.

**Prediction Time** Temps nécessaire au modèle pour prédire de nouvelles valeurs (125 dans notre cas).

**RF** Random Forest, Forêt aléatoire.

**RFSS** Risk Factor Scenarios Set.

**RMSE** Root Mean Square error, Racine de l'erreur quadratique moyenne.

**SCR** Solvency Capital Requirement.

**SdS** Simulations dans les Simulations.

**SVR** Support Vector Regression.

**Tuning Time** Temps nécessaire pour choisir les hyper-paramètres d'un modèle.

**WRE** Weighted Relative Erreur, erreur relative moyenne pondérée par la valeur de la variable cible.

**XGB** Extrême Gradient Boosting.

# Table des figures

1	Schématisation de l'approche des Simulations dans les Simulations <i>Source : Cours de modélisation et agrégation des risques</i> . . . . .	vi
2	Schématisation de l'approche Least Square Monte Carlo <i>Source : Documentation interne à Generali sur la méthodologie du modèle interne</i> . . . . .	vii
3	Métriques calculées sur la base de validation pour chaque modèle. . . . .	viii
4	Comparaison des SCR en fonction des fonctions répliquatives du BE utilisées	ix
5	Distribution des pertes de Fonds Propres stressés. En rouge est représentée la distribution par le polynôme et en bleu par MARS. . . . .	ix
6	Réaction des proxys en dehors des régions d'apprentissage de IdxBond . . . . .	x
7	Schema of the nested stochastic approach <i>Source : Risk modelling and aggregation course</i> . . . . .	xiii
8	Schematic representation of the Least Square Monte Carlo approach <i>Source : Generali internal documentation on the methodology of the internal model</i> . . . . .	xiv
9	Metrics calculated on the basis of validation for each model . . . . .	xv
10	Comparison of SCRs according to the replicative functions of the EB used	xvi
11	Distribution of the Losses of Stranded Equity In red is represented the polynomial distribution and in blue by MARS . . . . .	xvi
12	Reaction of proxies outside the IdxBond learning regions . . . . .	xvii
1.1	Bilan prudentiel simplifié . . . . .	4
1.2	Agrégation des risques en formule standard . . . . .	10
2.1	Schématisation de l'approche des Simulations dans les Simulations . . . . .	14
2.2	Schématisation de la localisation du scénario primaire quantile <i>Source : Cours de modélisation et d'agrégation des risques, M2 EURIA</i> . . . . .	17
2.3	Schématisation de l'approche <i>Curve Fitting</i> . <i>Source : Documentation interne à Generali sur la méthodologie du modèle interne</i> . . . . .	20
2.4	Schématisation de l'approche <i>Least Square Monte Carlo</i> <i>Source : Documentation interne à Generali sur la méthodologie du modèle interne</i> . . . . .	21

2.5	Différentes déformations de la courbe des taux prises en compte comme facteurs de risque . . . . .	23
3.1	Impact de chacun des facteurs de risque majeurs sur le BE individuellement	39
3.2	Impact de chacun des facteurs risques restants sur le BE individuellement	40
3.3	Relation liant les facteurs de risque au BE individuellement . . . . .	41
3.4	Relation liant les facteurs de risque au BE individuellement bis . . . . .	42
3.5	Importance des variables dans le modèle de forêt aléatoire dans l'explication de BE de la LoB unique. . . . .	46
3.6	Importance des variables dans le modèle XGBoost dans la régression du BE de la LoB unique. . . . .	49
3.7	Schéma du fonctionnement du modèle MARS . . . . .	51
3.8	Exemple illustratif d'un cas simple de SVR (Source : <a href="https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2">https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2</a> ) . . . . .	54
3.9	Exemple illustratif de l'astuce du noyau lors d'un SVR . . . . .	55
3.10	Schématisation de la validation croisée par k-fold Source : Wikipedia ( <a href="https://en.wikipedia.org/wiki/Cross-validation_(statistics)">https://en.wikipedia.org/wiki/Cross-validation_(statistics)</a> )	56
3.11	Valeurs prédites en fonction des valeurs de validation réelles sur la LoB unique par facteur de risque . . . . .	58
3.12	Valeurs prédites en fonction des valeurs de validation réelles par XGB . . . . .	59
3.13	Valeurs prédites en fonction des valeurs de validation réelles par MARS . . . . .	60
3.14	Valeurs prédites en fonction des valeurs de validation réelles par SVR . . . . .	61
3.15	Métriques calculées sur la base de validation pour chaque modèle. . . . .	63
3.16	Temps de construction et de prédiction des modèles . . . . .	63
3.17	Résidus en fonction de la valeur du <i>Best Estimate</i> . . . . .	64
4.1	Allure des proxies en 2D, pour chaque facteur de risque . . . . .	69
4.2	Allure des proxies en 2D, pour chaque facteur de risque . . . . .	70
4.3	Allure des proxies en 2D, pour chaque facteur de risque . . . . .	71
4.4	Comparaison des SCR en fonction des fonctions répliquatives du BE utilisées	74
4.5	Distribution des pertes de Fonds Propres stressés. En rouge est représentée la distribution par le polynôme et en bleu par MARS. . . . .	75
4.6	Étude de la distribution des pertes de Fonds Propres en fonction de l'approche utilisée . . . . .	75
4.7	Proportion des scénarios Out of Range sur les différentes variables explicatives . . . . .	77
4.8	Proportion des différents niveaux parmi les scénarios out of range. . . . .	78
4.9	Pertes de Fonds Propres obtenues sur le scénario exemple en fonction de la méthode . . . . .	79
4.10	Réaction des proxies en dehors des régions d'apprentissage de IdxBond . . . . .	80
4.11	Réaction des proxies en dehors des régions d'apprentissage de IdxEq . . . . .	81

D.1	Résultats obtenus sur la base de validation de la LoB Protection . . . . .	92
D.2	Résultats obtenus sur la base de validation de la LoB <i>pension</i> . . . . .	93
D.3	Résultats obtenus sur la base de validation de la LoB <i>pension</i> . . . . .	94
D.4	Résultats obtenus sur la base de validation de la LoB des unités de compte	96

# Bibliographie

- [1] Marsplines. <https://www.statsoft.fr/concepts-statistiques/multivariate-adaptive-regression-splines/MARSplines.php>.
- [2] Règlement délégué (UE) 2015/35 de la commission du 10 octobre 2014 complétant la directive 2009/138/ce du parlement européen et du conseil sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (solvabilité 2). *Journal officiel de l'Union Européenne*. (Texte présentant de l'intérêt pour l'EEE).
- [3] L. BREIMAN. Random forests. machine learning. 2001.
- [4] Laurent DEVINEAU and Mohammed BENNOUNA. Modélisation et agrégation des risques. *Support de cours M2 EURIA*, 2020-2021.
- [5] Laurent DEVINEAU and Stephane LOISEL. Construction d'un algorithme d'accélération de la méthode des "simulations dans les simulations" pour le calcul du capital économique solvabilité ii. *Bulletin Français d'Actuariat*, 2009.
- [6] DRUCKER and al. Support vector regression machines. 1996.
- [7] EUR-Lex. Directive 2009/138/ce du parlement européen et du conseil du 25 novembre 2009 sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (solvabilité ii). (Texte présentant de l'intérêt pour l'EEE).
- [8] R. GAUVILLE. Projection du ratio de solvabilité : des méthodes de machine learning pour contourner les contraintes opérationnelles de la méthode des sds. *Mémoire Institut des Actuaire*s, 2016.
- [9] R. GENUER and J-M POGGI. Arbres cart et forêts aléatoires, importance et sélection de variables. 2017.
- [10] Antonio Sutera Gilles Louppe, Louis Wehenkel and Pierre Geurts. Understanding variable importances in forests of randomized trees supplementary materials. 2013.
- [11] B. BOEHMKE & B. GREENWELL. Hands-on machine learning with r. 2020.
- [12] HASTIE, FRIEDMAN, and TIBSHIRANI. The element of statistical learning, data mining, inference and prediction.

- 
- [13] HOROWITZ and BADOURALY-KASSIM. Contexte règlementaire. *Support de cours M2 EURIA*, 2020-2021.
- [14] P. HOUNKONNOU. Modélisation alm et estimation efficace du scr d'un assureur vie en modèle interne avec le krigeage stochastique. *Mémoire Institut des Actuaire*s, 2020.
- [15] E. KARTAL KOC and H. BOZDOGAN. Model selection in multivariate adaptive regression splines (mars) using information complexity as the fitness function. 2014.
- [16] LIU and STAUM. Stochastic kriging for efficient nested simulation of expected shortfall. 2017.
- [17] A. MAZIERE. Cartographie de l'apprentissage artificiel et de ses algorithmes. *Thèse de doctorat*, 2021.
- [18] Stephen MILBORROW. Notes on the earth package. 2021.
- [19] J. PARR and T. HOWARD. Gradient boosting performs gradient descent. <https://explained.ai/gradient-boosting/descent.html>.
- [20] F. ROUCHATI. Projection du ratio de solvabilité d'un assureur retraite : les méthodes paramétriques ( curve fitting et least squares monte carlo) peuvent-elles se substituer à la méthode de simulations dans les simulation ? *Mémoire Institut des Actuaire*s, 2016.
- [21] L. ROUVIÈRE. Machine learning. 2021.
- [22] G. SAINT-CIRGUE. Descente de gradient - gradient descent. 2019. <https://machinelearnia.com/descente-de-gradient/>.
- [23] L. STENTOFT. Convergence of the least square monte carlo approach to american option valuation. *Management Science*, 50, 2004.
- [24] Franck VERMET. Apprentissage statistique : une approche connexionniste. *Support de cours M1 EURIA*, 2019-2020.
- [25] P. VOUZNI. Calcul de la probability distribution forecast du best estimate via l'approche lsmc à partir d'un modèle de flexing. *Mémoire Institut des Actuaire*s, 2017.