

**Mémoire présenté devant le CNAM pour l'obtention du Master
Droit Economie Gestion, mention actuariat
et l'admission à l'Institut des Actuaires
23 mars 2022**

Par : Botrel KALAWÉ KALAWÉ

Titre : CALIBRAGE D'UN MODELE D'AGREGATION DE DONNÉES POUR LA
PRODUCTION DE COMPTES IFRS17 EN METHODE VFA

Confidentialité : Non

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

Président du Jury :

M. Nathanaël ABECERA

Entreprise :

Allianz France

Directeur de mémoire en entreprise

Membres présents du jury de l'institut
des Actuaires :

M. Michel GERMAIN

M. Sylvain CARACO

Nicolas BOURE

Signature :

Invités :

O. RENAUDIN, E. CODET

Membres présents du jury du CNAM :

M. Olivier DESMETTRE

M. David FAURE

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 en entreprise :

Secrétariat :

Signature du candidat :

Bibliothèque :

REMERCIEMENTS

Je tiens à remercier mon Directeur de mémoire, Nicolas BOURE pour sa disponibilité et ses remarques pertinentes.

Merci à toute l'équipe pédagogique du master d'actuariat du CNAM. Mon encadreur Olivier DESMETTRE, pour ses remarques, et à Michel FROMENTEAU pour la qualité de la formation.

Merci à mes collègues pour leurs conseils et en particulier à Georges pour ses astuces sur Python.

Je souhaiterais enfin remercier mes proches (parents, frères et sœurs et amis) pour leur encouragement, et en particulier mon épouse qui a su m'apporter son soutien tout au long de ces années de formation.

RESUME

La norme IFRS 17 qui entre en vigueur en janvier 2023 va profondément modifier la communication financière des compagnies d'assurance. IFRS 17, qui remplace IFRS4, impose d'évaluer et de communiquer les profits futurs attendus sur les contrats d'assurance comportant une composante de risques. Elle exige en outre une production de l'information financière à une granularité plus fine.

Le changement de l'outil de projection des cash flows et de modélisation ALM (Asset and Liability Management), passage d'un modèle ALM *Flexing* à un modèle *Full* ALM permettra de répondre en partie à cette exigence. Avec un modèle *Full* ALM la modélisation est effectuée à la maille *model point* plus fine que la maille poche de passifs utilisée dans un modèle ALM *Flexing*. Pour avoir des temps de traitement acceptables, l'utilisation du modèle *Full* ALM nécessite cependant une agrégation préalable de ces *model points*.

Ce mémoire a pour objectif d'améliorer le calibrage de l'algorithme actuellement utilisé pour agréger les *model points*. Un calibrage qui se fait d'une part au niveau de la fonction cible et d'autre part au niveau des données en entrée de l'algorithme. Nous effectuerons notre étude sur un portefeuille de prévoyance individuelle.

Après avoir rappelé les exigences introduites par IFRS 17 et les évolutions de l'outil de projection, nous procéderons à une description des données de notre portefeuille. Une analyse du modèle d'agrégation nous permettra par la suite d'identifier ces leviers d'optimisation. Grâce à l'étude des corrélations, la combinaison d'un algorithme de classification non supervisée et l'analyse en composantes principales, nous proposerons un calibrage de ce modèle.

Mots clés : Moindres Carrés Non Négatifs, Clustering (K-MEANS), Analyse en Composantes Principales, BEL, *Model Points*

ABSTRACT

IFRS 17 which will be applied in January 2023 will profoundly change the financial reporting of insurance companies. IFRS 17 which will replace IFRS 4, requires measuring and disclosing expected future profits of these contracts at a more detailed granularity. Changing the modeling tool (cash flows projection) from an ALM (Asset and Liability Management) Flexing model to a Full ALM model will help to meet these requirements. With a Full ALM model, modeling is performed using a finer model points than the “Liability” used in an ALM Flexing model. To have acceptable run times, however, the use of the Full ALM model requires prior aggregation of these model points.

This thesis aims to improve the calibration of the algorithm currently used to aggregate the model points on individual protection scope. A calibration that is done firstly at the level of the target function and secondly at the level of the input data of the algorithm.

After restating the requirements introduced by IFRS 17 and the change made on the modeling tool, we will describe the portfolio. An analysis of the aggregation model will then allow us to identify these optimization levers. Through the study of correlations, the combination of an unsupervised classification algorithm and principal component analysis, we will propose a calibration of the model.

Keywords: Non-Negative Least Squares, Clustering, Principal Component Analysis, Best Estimate Liability, Model Points.

NOTE DE SYNTHÈSE

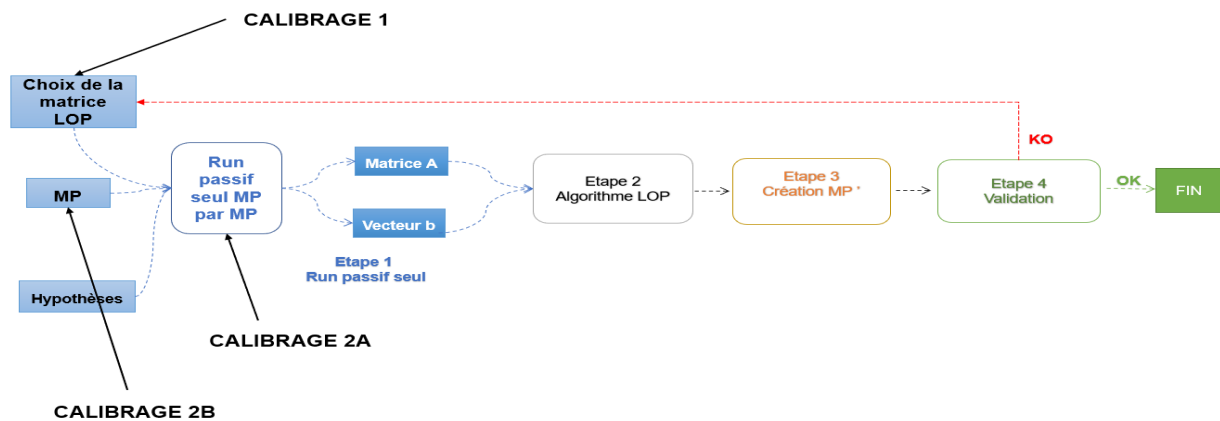
CONTEXTE ET PROBLÉMATIQUE

La norme IFRS 17 qui entre en vigueur en janvier 2023 va profondément modifier la communication financière des compagnies d'assurance. IFRS 17 impose d'évaluer et de communiquer les profits futurs attendus sur les contrats d'assurance ayant une composante de risques. Elle exige en outre une production de l'information financière à une granularité plus fine. Le changement de l'outil de projection des cash flows et de la modélisation ALM (Asset and Liability Management), passage d'un modèle ALM *Flexing* à un modèle *Full* ALM permettra de répondre en partie à cette exigence. Le modèle *Flexing*, associant modèles déterministe et stochastique, est remplacé par un modèle *Full* ALM de modélisation directe des interactions entre actif et passif afin d'obtenir des projections de cash flows à une maille plus fine. Pour avoir des temps de traitement acceptables, l'utilisation du modèle *Full* ALM nécessite cependant une agrégation préalable de ces model points.

Ce mémoire a pour objectif d'améliorer le calibrage de l'algorithme, *Linear Optimisation Approach* (LOP), actuellement utilisé pour agréger les *model points*, ceci sans modifier les processus de création des MP et sans modifier l'algorithme LOP. Nous nous intéresserons au périmètre prévoyance individuelle, périmètre sur lequel la qualité d'agrégation n'était pas satisfaisante durant les premières clôtures à blanc des comptes IFRS 17.

MÉTHODE SUIVIE ET RESULTATS

L'analyse du modèle d'agrégation actuellement utilisé nous a permis d'identifier ses leviers d'optimisation, leviers qui nous permettront de proposer un calibrage de ce modèle d'agrégation. La figure suivante présente le processus d'agrégation en mettant en évidence les trois axes identifiés.



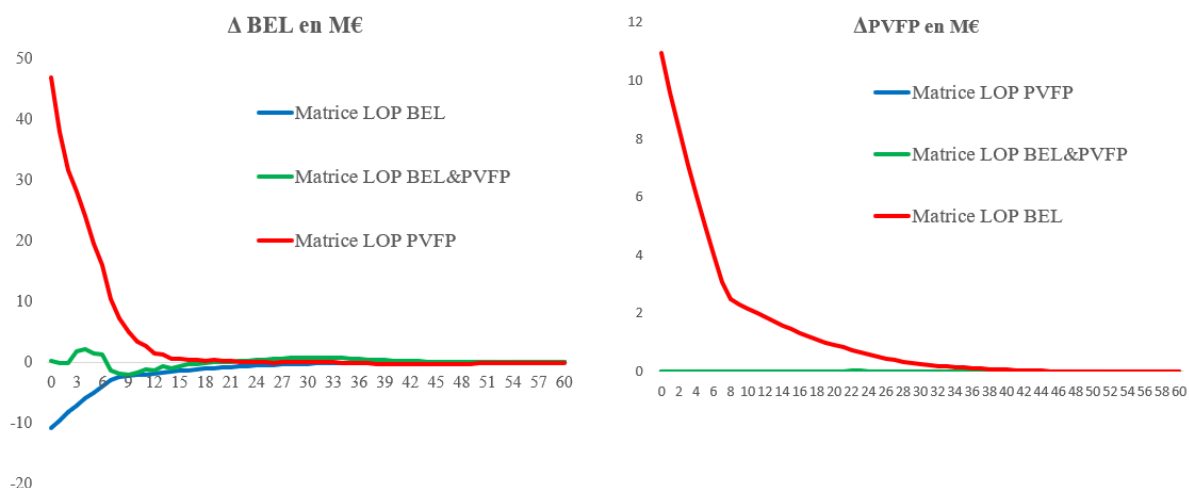
L'algorithme LOP basé sur les moindres carrés non négatifs (NNLS), consiste à résoudre l'équation $Ax = b$ avec $x \geq 0$. Pour optimiser cette équation, il faut optimiser la matrice A et le vecteur b, c'est-à-dire les inputs de l'algorithme. Optimiser les inputs revient à bien choisir la matrice LOP (**calibrage 1**) et à bien calibrer le run passif seul MP après MP. La calibration du run passif seul, peut se faire soit directement sur ces sorties (**calibrage 2A**) soit sur les inputs de ce run, les MP (**calibrage 2B**). En effet, les MP inputs de ce run sont eux-mêmes pré-regroupés ; les hypothèses ne sont pas modifiées durant le processus.

1. Calibrage de la matrice LOP (calibrage 1)

Nous avons identifié les variables qui sont les plus corrélées à différents indicateurs (BEL, PVFP, BEL et PVFP). Ces variables nous permettent de construire trois matrices, matrices que nous utiliserons par la suite pour exécuter l'algorithme.

a. Analyse sur toutes les années de projection

Les figures ci-dessous comparent le BEL et la PVFP avant agrégation et le BEL et la PVFP après agrégation selon que la matrice LOP est issue de la corrélation avec le BEL, la PVFP ou le BEL et la PVFP.



Jusqu'à $t = 6$, le BEL basé sur la matrice LOP PVFP s'écarte du BEL avant agrégation. Les autres BEL (matrice LOP BEL, matrice LOP BEL & PVFP) sont à peu près identiques au BEL avant agrégation sur toute la durée de projection.

Les écarts de PVFP calculés entre la matrice LOP avant agrégation et les matrices LOP BEL & PVFP d'une part et la matrice LOP PVFP d'autre part sont nuls sur toute la durée de projection.

L'écart entre la PVFP calculée grâce à la matrice LOP BEL s'écarte de la PVFP avant agrégation de plus d'un million d'euros jusqu'à $t = 8$. Après cette année, l'écart est inférieur à un million d'euros et devient nul autour de $t = 35$.

b. Analyse à $t=0$

Les résultats de cette étape du calibrage (à $t=0$) sont regroupés dans le tableau ci-après. ϵ_{Bel} et ϵ_{PVFP} représentent respectivement les écarts relatifs par rapport au BEL et à la PVFP avant agrégation

Technique	Critères	BEL (€)	ϵ_{Bel}	PVFP (€)	ϵ_{PVFP}	MP
Données avant agrégation		818 096 469		124 417 575		18 525
Corrélation (Pearson)	BEL	807 222 306	1%	135 333 677	9%	4 015
	PVFP	865 072 070	6%	124 417 720	0%	4 274
	BEL & PVFP	818 370 729	0%	124 421 328	0%	4 230

Synthèse des vecteurs LOP par critères et impacts BEL et PVFP

Les meilleurs résultats sont ceux obtenus lorsque les variables utilisées sont celles issues de la corrélation entre le BEL et la PVFP. En revanche, avec un taux de compression de 77% (1-4230/18 525), le nombre de MP après agrégation reste élevé. Nous devons trouver d'autres pistes d'agrégation, cette fois basées sur les critères de regroupement des données.

2. Calibrage en entrées du modèle

a. Pré-Clustering sur les cash-flows de passif (calibrage 2A)

Nous appliquons à cette étape l'algorithme du K-Means sur les cash flows issus du run passif seul MP par MP. Nous avons retenu 3 indicateurs de classification qui sont représentatifs du bilan d'une compagnie d'assurance vie : le BEL, la PM et la PVFP. Dans les 7 configurations possibles, le nombre de clusters optimal est de 15 et la répartition des données dans les clusters se présentent comme suit :

Cluster	BEL	PM	PVFP	BEL &PVFP	PVFP &PM	BEL &PM	BEL &PM&PVFP
Cluster 1	12 649	11 188	16 307	15 837	13 025	13 678	15 085
Cluster 2	3 143	3 518	825	1 517	2 974	2 754	1 998
Cluster 3	1 235	1 623	574	381	1 139	1 047	576
Cluster 4	575	897	288	311	501	441	261
Cluster 5	296	475	152	136	263	248	259
Cluster 6	213	286	126	93	201	112	114
Cluster 7	120	172	77	71	131	80	80
Cluster 8	98	123	46	42	95	65	39
Cluster 9	84	82	45	42	84	31	31
Cluster 10	40	80	29	33	40	21	28
Cluster 11	28	40	22	23	28	18	17
Cluster 12	24	21	11	14	24	11	12
Cluster 13	11	11	10	11	11	10	12
Cluster 14	6	6	8	11	6	6	10
Cluster 15	3	3	5	3	3	3	3
Total	18 525	18 525	18 525	18 525	18 525	18 525	18 525

Clusters basés sur les variables run MP / MP

Ces résultats nous montrent que la majorité des données sont concentrées dans le premier cluster (plus de 60% des données) dans tous les cas. En outre, les 5 derniers clusters (cluster 15, cluster 14, cluster 13, cluster 12 et cluster 11) ont quasiment la même taille dans les 7 cas. On observe toutefois une exception lorsque l'indicateur de lancement de l'algorithme est la PM. Dans ce dernier cas, le cluster 11 a deux fois plus de données que dans les autres cas.

Nous allons exécuter l'algorithme LOP dans chacun de ces 7 cas. Les variables LOP en input étant celles que nous avons obtenues à la première étape du calibrage. Le tableau ci-après présente les résultats obtenus à $t=0$.

Technique	Critères	BEL (€)	ε Bel	PVFP (€)	ε PVFP	MP
Données avant agrégation		818 096 469		124 417 575		18 525
Clustering (K-Means)	BEL	579 978 307	-29%	124 335 320	-0,1%	576
	PM	639 839 004	-22%	124 407 947	0,0%	610
	PVFP	539 311 370	-34%	124 376 114	0,0%	492
	BEL &PVFP	519 329 738	-37%	124 230 556	-0,2%	473
	PVFP &PM	512 349 252	-37%	124 273 607	-0,1%	574
	BEL &PM	737 028 768	-10%	124 403 306	0,0%	500
	BEL &PM& PVFP	892 832 540	9%	124 543 663	0,1%	494

Agrégation sur la base du Clustering K-Means

Ces résultats montrent tout d'abord que l'écart relatif par rapport à la PVFP reste acceptable dans tous les cas. En outre, avec le clustering, nous arrivons à réduire considérablement le nombre de MP par rapport aux résultats obtenus dans le calibrage 1. Cependant l'écart relatif par rapport au BEL se dégrade. Ceci probablement à cause du cluster 1, comme nous l'avons indiqué, qui représente plus de 60% des données. En résumé, le calibrage 1 nous permet d'avoir des écarts relatifs acceptables avec une mauvaise qualité d'agrégation. Le calibrage 2A cependant nous propose exactement l'inverse.

b. Pré-Clustering sur les composantes principales des MP (calibrage 2B)

En nous fondant sur un théorème énoncé par Ding & He (2004)¹, nous appliquons le K-Means sur les composantes principales des MP. Comme précédemment, le nombre optimal de cluster est toujours 15. Nous obtenons la répartition suivante des données :

¹ Le sous-espace formé par les centroïdes de la solution du clusters K-Means est incorporé dans les premières K-1 composantes principales.

Cluster	Première composante	2 premières composantes	3 premières composantes	4 premières composantes
Cluster 1	2 018	3 630	4 472	16 686
Cluster 2	2 005	3 255	4 059	757
Cluster 3	1 999	3 167	3 156	218
Cluster 4	1 716	2 700	2 830	155
Cluster 5	1 703	2 063	983	142
Cluster 6	1 689	974	843	133
Cluster 7	1 606	710	611	131
Cluster 8	1 421	639	573	131
Cluster 9	1 296	612	467	83
Cluster 10	1 026	383	334	44
Cluster 11	563	218	126	22
Cluster 12	531	108	50	15
Cluster 13	442	49	17	5
Cluster 14	392	14	3	2
Cluster 15	118	3	1	1
Total	18 525	18 525	18 525	18 525

Clusters basés sur ACP sur les MP

De cette répartition, on observe d'une part que le cluster 1 comporte moins de 25% des données dans tous les cas, sauf quand le regroupement est effectué sur la base des 4 premières composantes. D'autre part, les données dans le cas d'un regroupement sur la base de la première composante principale sont très uniformes et peu dispersées : plus on ajoute des composantes, moins la répartition des données est uniforme et plus les données sont dispersées entre les clusters. Par ailleurs, les dix premiers clusters dans cette dernière configuration contiennent entre 2 018 et 1 026 données. Le dernier cluster regroupe plus de 100 individus.

Nous allons maintenant exécuter l'algorithme LOP sur ces données. Nous obtenons les résultats suivants à $t=0$:

Technique	Critères	BEL (€)	$\underline{\epsilon}$ Bel	PVFP (€)	$\underline{\epsilon}$ PVFP	MP
Données avant agrégation		818 096 469		124 417 575		18 525
Clustering (K-Means) + ACP	Première composante	789 955 751	-3%	124 400 301	0%	903
	2 premières composantes	663 749 376	-19%	124 331 376	0%	738
	3 premières composantes	624 027 512	-24%	124 359 304	0%	691
	4 premières composantes	417 909 577	-49%	124 269 096	0%	617

Agrégation sur la base du Clustering K-Means après ACP

Le résultat obtenu après regroupement sur la base de la première composante donne le meilleur compromis erreur BEL, erreur PVFP et compression des MP. Ce résultat combine les avantages du calibrage 1 et du calibrage 2A. Nous proposons de retenir cette configuration pour le calibrage de notre modèle.

INTERPRETATION ET CONCLUSION

Notre étude permet de montrer ainsi qu'en combinant les moindres carrés non négatifs (NNLS) avec un algorithme de clustering on arrive à des meilleurs résultats qu'en utilisant uniquement les NNLS. Par ailleurs, en effectuant le clustering sur les MP on obtient des meilleurs résultats qu'en les effectuant sur les variables de cash-flows déterministes basées sur ces MP. Une application de la démarche de calibration proposée dans cette étude sur d'autres périmètres permettra de généraliser ces conclusions. Notons aussi que notre étude a été limitée à une analyse des sorties déterministes. Nous n'avons pas effectué une analyse stochastique, notamment des temps de run stochastiques. Mais a priori plus le nombre de MP est réduit, plus le temps de run sera court, toutes choses étant égales par ailleurs.

Pour finir, les résultats de cette étude suscitent une question sur la constitution même des MP. En effet plutôt que remplacer l'algorithme LOP (NNLS), ne faudrait-il pas plutôt repenser le processus de création des MP pour respecter les contraintes d'agrégation qu'impose un modèle full ALM ? Dans ce cas, le pré-clustering s'effectuerait sur la base des contrats plutôt que sur les MP.

SYNTHESIS NOTE

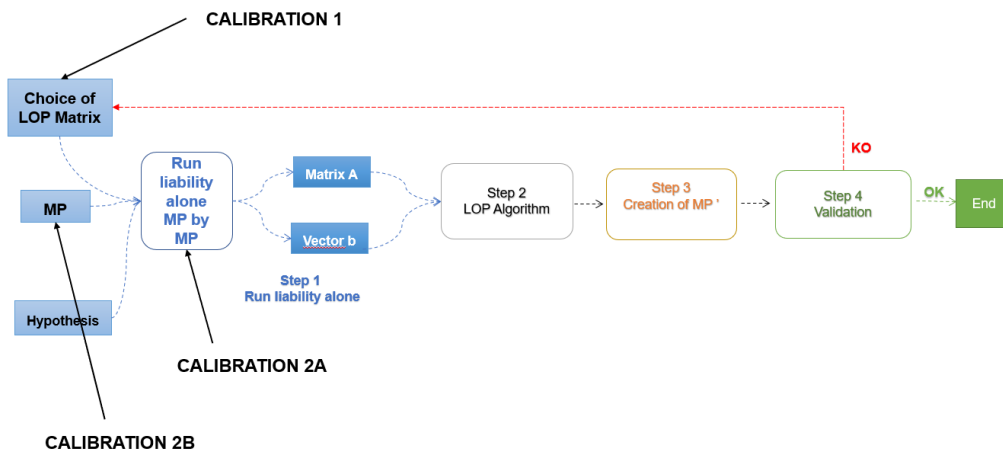
BACKGROUND AND ISSUES

IFRS 17, will be applied in January 2023, will profoundly change the financial communication of insurance companies. IFRS 17 requires the measurement and disclosure of expected future profits on insurance contracts with a risk component. It also requires the production of financial information at a finer granularity. The change of the cash flows projection tool and the ALM (Asset and Liability Management) modeling, from an ALM Flexing model to a Full ALM model, will help to meet these requirements. The Flexing model, combining deterministic and stochastic models, is replaced by a Full ALM model of direct modeling of interactions between assets and liabilities in order to obtain cash flows projections at a finer scale. In order to have acceptable run times, the use of the Full ALM model requires a prior aggregation of these model points.

The objective of this thesis is to improve the calibration of the Linear Optimisation Approach (LOP) algorithm currently used to aggregate model points (MP), without modifying the MP creation processes and without modifying the LOP algorithm. We will focus on the individual pension perimeter, where the quality of aggregation was not satisfactory during the first IFRS 17 dry runs.

METHOD USED AND RESULTS

The analysis of the aggregation model currently used has allowed us to identify its optimization levers, levers that will allow us to propose a calibration of this aggregation model. The following figure presents the aggregation process by highlighting the three identified axes



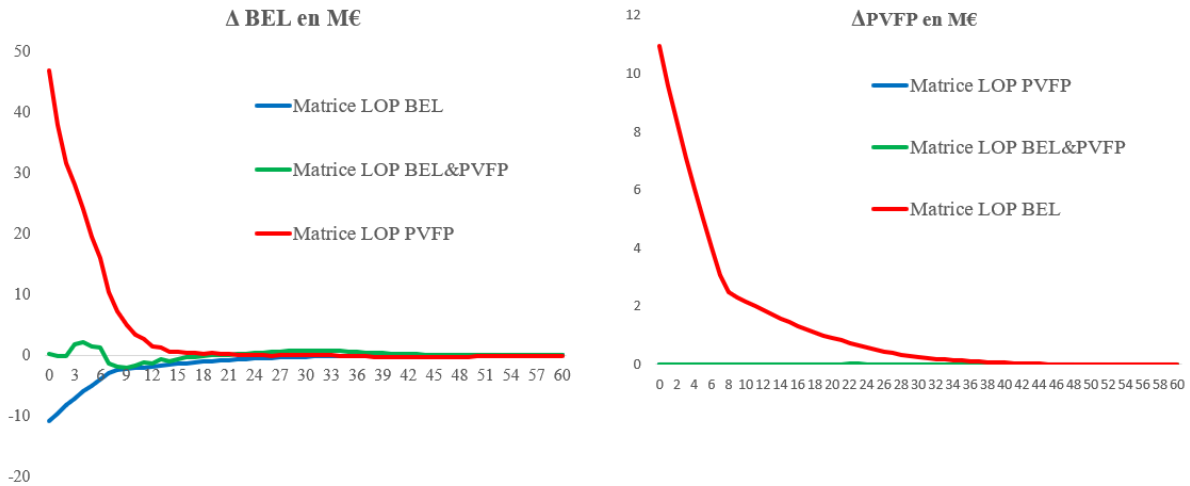
The LOP algorithm based on non-negative least squares (NNLS), consists of solving the equation $Ax = b$ with $x \geq 0$. To optimize this equation, we need to optimize the matrix A and the vector b, i.e., the inputs of the algorithm. Optimizing the inputs amounts to choosing the LOP matrix well (calibration 1) and calibrating the run liability alone MP by MP well. The calibration of this run can be done either directly on these outputs (calibration 2A) or on the inputs of the run, the MP (calibration 2B). Indeed, the MP inputs of this run are themselves pre-grouped ; the assumptions are not modified during the process.

1. Calibration of the LOP's matrix (calibration 1)

We have identified the variables that are most correlated to different indicators (BEL, PVFP, BEL and PVFP). These variables allow us to build three matrices, matrices that we will use later to run the algorithm.

a. Analysis over all projection years

The figures below compare BEL and PVFP before aggregation and BEL and PVFP after aggregation depending on whether the LOP matrix is derived from the correlation with BEL, PVFP or BEL and PVFP.



Until $t = 6$, the BEL based on the LOP PVFP matrix deviates from the pre-aggregation BEL. The other BELs (LOP BEL matrix, LOP BEL & PVFP matrix) are approximately the same as the pre-aggregation BEL throughout the projection period.

The calculated PVFP deviations between the pre-aggregation LOP matrix and the LOP BEL & PVFP matrices on the one hand and the LOP PVFP matrix on the other hand are zero over the entire projection period.

The difference between the PVFP calculated with the LOP BEL matrix deviates from the PVFP before aggregation by more than one million euros until $t = 8$. After this year, the difference is less than one million euros and becomes zero around $t = 35$.

b. Analysis at $t=0$

The results of this stage of the calibration (at $t=0$) are grouped in the table below. ε_{Bel} and ε_{PVFP} represent the relative deviations from BEL and PVFP respectively before aggregation.

Approach	Criteria	BEL (€)	ε_{Bel}	PVFP (€)	ε_{PVFP}	MP
Data before aggregation		818 096 469		124 417 575		18 525
Correlation (Pearson)	BEL	807 222 306	1%	135 333 677	9%	4 015
	PVFP	865 072 070	6%	124 417 720	0%	4 274
	BEL & PVFP	818 370 729	0%	124 421 328	0%	4 230

Summary of LOP vectors by criteria and BEL and PVFP impacts

The best results are those obtained when the variables used are those from the correlation between BEL and PVFP. On the other hand, with a compression ratio of 77% ($1 - 4\,230/18\,525$), the number of MP after aggregation remains high. We need to find other avenues of aggregation, this time based on the clustering criteria of the data.

2. Calibration of inputs

a. Pre-Clustering based on cash flows of liabilities (calibration 2A)

At this stage, we apply the K-Means algorithm on the cash flows from run liability alone MP by MP well. We have selected 3 classification indicators that are representative of the balance sheet of a life insurance company: BEL, PM and PVFP. In the 7 possible configurations, the optimal number of clusters is 15 and the distribution of data in the clusters is as follows :

Cluster	BEL	PM	PVFP	BEL &PVFP	PVFP &PM	BEL &PM	BEL &PM& PVFP
Cluster 1	12 649	11 188	16 307	15 837	13 025	13 678	15 085
Cluster 2	3 143	3 518	825	1 517	2 974	2 754	1 998
Cluster 3	1 235	1 623	574	381	1 139	1 047	576
Cluster 4	575	897	288	311	501	441	261
Cluster 5	296	475	152	136	263	248	259
Cluster 6	213	286	126	93	201	112	114
Cluster 7	120	172	77	71	131	80	80
Cluster 8	98	123	46	42	95	65	39
Cluster 9	84	82	45	42	84	31	31
Cluster 10	40	80	29	33	40	21	28
Cluster 11	28	40	22	23	28	18	17
Cluster 12	24	21	11	14	24	11	12
Cluster 13	11	11	10	11	11	10	12
Cluster 14	6	6	8	11	6	6	10
Cluster 15	3	3	5	3	3	3	3
Total	18 525	18 525	18 525	18 525	18 525	18 525	18 525

Clusters based on run MP / MP variables

These results show us that the majority of the data is concentrated in the first cluster (more than 60% of the data) in all cases. Moreover, the last 5 clusters (cluster 15, cluster 14, cluster 13, cluster 12 and cluster 11) have almost the same size in all 7 cases. However, we observe an exception when the algorithm's start indicator is PM. In the latter case, cluster 11 has twice as much data as in the other cases.

We will run the LOP algorithm in each of these 7 cases. The LOP variables in input being those we obtained in the first step of the calibration. The table below shows the results obtained at $t=0$

Approach	Criteria	BEL (€)	€ Bel	PVFP (€)	€ PVFP	MP
Data before aggregation		818 096 469		124 417 575		18 525
Clustering (K-Means)	BEL	579 978 307	-29%	124 335 320	-0,1%	576
	PM	639 839 004	-22%	124 407 947	0,0%	610
	PVFP	539 311 370	-34%	124 376 114	0,0%	492
	BEL &PVFP	519 329 738	-37%	124 230 556	-0,2%	473
	PVFP &PM	512 349 252	-37%	124 273 607	-0,1%	574
	BEL &PM	737 028 768	-10%	124 403 306	0,0%	500
	BEL &PM& PVFP	892 832 540	9%	124 543 663	0,1%	494

Aggregation based on K-Means clustering

s

These results first show that the relative deviation from PVFP remains acceptable in all cases. Moreover, with clustering, we manage to reduce the number of MPs considerably compared to the results obtained in calibration 1. However, the relative deviation from BEL deteriorates. This is probably due to cluster 1, as we have indicated, which represents more than 60% of the data. In summary, calibration 1 allows us to have acceptable relative deviations with a poor aggregation quality. Calibration 2A, however, offers us exactly the opposite.

b. Pre-Clustering on principal components of MP (Calibration 2B)

Based on a theorem stated by Ding & He (2004)², we apply K-Means on the principal components of MP. As before, the optimal number of clusters is always 15. We obtain the following distribution of the data :

Cluster	First component	First 2 components	First 3 components	First 4 components
Cluster 1	2 018	3 630	4 472	16 686
Cluster 2	2 005	3 255	4 059	757
Cluster 3	1 999	3 167	3 156	218
Cluster 4	1 716	2 700	2 830	155
Cluster 5	1 703	2 063	983	142
Cluster 6	1 689	974	843	133
Cluster 7	1 606	710	611	131
Cluster 8	1 421	639	573	131
Cluster 9	1 296	612	467	83
Cluster 10	1 026	383	334	44
Cluster 11	563	218	126	22
Cluster 12	531	108	50	15
Cluster 13	442	49	17	5
Cluster 14	392	14	3	2
Cluster 15	118	3	1	1
Total	18 525	18 525	18 525	18 525

Clusters based on PCA on the MP

From this distribution, we observe on the one hand that cluster 1 comprises less than 25% of the data in all cases, except when the grouping is made on the basis of the first 4 components. On the other hand, the data in the case of a grouping based on the first principal component are very uniform and not very dispersed: the more components are added, the less uniform the distribution of data is and the more dispersed the data are between the clusters. Moreover, the

² The subspace formed by the centroids of the K-Means cluster solution is incorporated into the first K- 1 principal components.

first ten clusters in this last configuration contain between 2,018 and 1,026 data. The last cluster contains more than 100 individuals.

We will now run the LOP algorithm on these data. We obtain the following results at t=0 :

Approach	Criteria	BEL (€)	ε Bel	PVFP (€)	ε PVFP	MP
Data before aggregation		818 096 469		124 417 575		18 525
Clustering (K-Means) + ACP	First component	789 955 751	-3%	124 400 301	0%	903
	First 2 components	663 749 376	-19%	124 331 376	0%	738
	First 3 components	624 027 512	-24%	124 359 304	0%	691
	First 4 components	417 909 577	-49%	124 269 096	0%	617

Aggregation based on K-Means clustering after PCA

The result obtained after clustering based on the first component gives the best compromise of BEL error, PVFP error and MP compression. This result combines the advantages of calibration 1 and calibration 2A. We propose to retain this configuration for the calibration of our model.

INTERPRETATION AND CONCLUSION

Our study shows that combining non-negative least squares (NNLS) with a clustering algorithm yields better results than using NNLS alone. Moreover, by performing clustering on the MP we obtain better results than by performing it on the deterministic cash flows variables based on these MP. An application of the calibration approach proposed in this study on other perimeters will allow to generalize these conclusions. Note also that our study was limited to an analysis of deterministic outputs. We did not perform a stochastic analysis, in particular of stochastic run times. But, *a priori*, the smaller the number of MP, the shorter the run time, all other things being equal.

Finally, the results of this study raise a question about the constitution of the MP themselves. Indeed, rather than replacing the LOP algorithm (NNLS), should we not rather rethink the MP creation process to respect the aggregation constraints imposed by a full ALM model and IFRS 17 ? In this case, pre-clustering would be based on contracts rather than on MP.

TABLE DE MATIERES

REMERCIEMENTS	i
RESUME	ii
ABSTRACT	iii
NOTE DE SYNTHÈSE	iv
SYNTHESIS NOTE	xi
TABLE DE MATIERES	xvii
INTRODUCTION	1
1. CONTEXTE DE L'ETUDE ET PROBLEMATIQUE	3
1.1 L'introduction de la norme IFRS 17	3
1.1.1 Présentation et principaux enjeux.....	3
1.1.1.1 Présentation générale de la norme	3
1.1.1.2 Exigences en termes de maille de calcul	8
1.1.1.3 Les 3 modèles d'évaluation possibles	10
1.2 L'évolution du modèle ALM.....	13
1.2.1 Les limites du modèle <i>flexing</i> actuel (ALIM)	13
1.2.1.1 Modèle déterministe (MoDet)	13
1.2.1.2 La nécessité d'un modèle <i>flexing</i>	14
1.2.1.3 Modélisation stochastique	16
1.2.2 La nécessité de passer à un modèle <i>Full ALM</i> (LSIM).....	18
1.2.2.1 Objectifs	18
1.3 Problématique de l'étude.....	20
2. PRESENTATION DES DONNEES	22
2.1 Prévoyance individuelle	22
2.1.1 Généralités.....	22
2.1.2 Types de garanties	22
2.1.3 Les produits du portefeuille.....	24
2.2 Présentation des bases de données	24
2.2.1 Les <i>model points</i>	25
2.2.1.1 Analyse univariée	25
2.2.1.2 Analyse bivariée	31
2.2.2 Base de données des cash flows déterministes.....	33
2.2.2.1 Contribution des garanties au BEL.....	35

2.2.2.2	Contribution des MP au BEL	35
3.	MODELE D'AGREGATION DES DONNEES	37
3.1	<i>Non-Negative Least Squares</i> (NNLS)	37
3.1.1	Moindres carrés	37
3.1.1.1	Théorème de Perron-Frobenius	38
3.1.1.2	Conditions de Karush-Kuhn-Tucker	40
3.1.2	La méthode des moindres carrés non négatifs.....	41
3.1.3	L'algorithme NNLS.....	43
3.2	L'algorithme NNLS chez Allianz	44
3.2.1	Approche méthodologique	44
3.2.2	Processus LOP dans le modèle LSIM	45
3.3	Applications sur le périmètre prévoyance individuelle	48
4.	CALIBRAGE DU MODELE	49
4.1	Calibrage de la matrice LOP	49
4.1.1	Calibrage par rapport au BEL.....	50
4.1.2	Calibrage par rapport à la PVFP.....	51
4.1.3	Calibrage par rapport au BEL et la PVFP	53
4.1.4	Analyse des résultats	53
4.2	Calibrage en entrée du modèle.	56
4.2.1	Pré-Clustering sur les cash-flows de passif	56
4.2.1.1	Critères d'homogénéité des groupes : le théorème de Huygens.....	57
4.2.1.2	K-MEANS.....	59
4.2.1.3	Mise en œuvre de l'algorithme du K-Means	60
4.2.2	Pré-Clustering sur les composantes principales des MP	64
4.2.2.1	Lien entre l'ACP et l'algorithme du K-Means	64
4.2.2.2	Mise en œuvre du K-Means sur les composantes principales	65
4.2.3	Analyse des résultats	84
	CONCLUSION.....	85
	ANNEXES.....	87
	Annexe 1 : Statistiques descriptives	87
	Annexe 2 : Choix de la matrice LOP.....	93
	Annexe 3 : Comparaison des indicateurs constituant le BEL.....	97
	Annexe 4 : Principe général de l'ACP.....	104
	BIBLIOGRAPHIE	108
	LISTE DES FIGURES.....	110
	LISTE DES TABLEAUX	111

INTRODUCTION

Sept ans après l'entrée en vigueur de Solvabilité 2 en janvier 2016, les compagnies d'assurance se préparent à l'entrée en vigueur en 2023 de la norme IFRS 17. Cette norme modifie profondément la communication financière. IFRS 17, qui remplace IFRS4, s'applique uniquement aux contrats comportant une composante de risques (technique ou financier) pour les sociétés d'assurance et impose d'évaluer et de communiquer les profits futurs attendus de ces contrats. Elle exige en outre une production de l'information financière à une granularité plus fine. Cette production détaillée de la performance financière risque d'être réalisée au départ dans un contexte économique particulier (taux bas, pandémie de la Covid-19, etc.). Ce contexte impose la transformation des directions financières : elles doivent revoir leurs processus et optimiser leurs outils de modélisation.

C'est dans ce contexte qu'un changement de l'outil de projection des cash flows et de modélisation ALM (Asset and Liability Management) a été initié. Le modèle *Flexing*, associant modèles déterministe et stochastique, est remplacé par un modèle *Full ALM*, de modélisation directe des interactions entre actif et passif afin d'obtenir des projections de cash flows à une maille plus fine, ceci pour répondre aux exigences d'IFRS 17. En effet, dans un modèle *Flexing*, les inputs sont les hypothèses et les poches de passifs. Sur le périmètre épargne, retraite et prévoyance longue, périmètre sur lequel s'applique principalement la méthode d'évaluation VFA (Variable Fee Approach) en IFRS 17, il y a environ 500 poches de passifs. Dans le modèle *Full ALM* en revanche, les inputs sont les hypothèses et les *model points*. Il y a environ 120 000 *model points* sur ce même périmètre. Pour avoir des temps de traitement acceptables en *Full ALM*, il faut agréger ces *model points*. Un algorithme, le *Linear Optimisation Approach* (LOP), a été mis en place pour effectuer cette agrégation.

Sur le périmètre prévoyance individuelle, durant les premières clôtures à blanc des comptes IFRS 17, la qualité d'agrégation n'était pas satisfaisante : avec 18 525 *model points* (MP) initiaux, l'algorithme permettait d'obtenir 4 056 MP agrégés, soit une réduction de 78%. Le BEL (Best Estimate Liability) initial était de 818 millions €, le BEL après agrégation de 787 millions €, soit une erreur d'estimation de 3,78%.

L'objectif de notre étude est d'optimiser la qualité d'agrégation sur ce dernier périmètre, ceci sans modifier les processus de création des MP et sans modifier l'algorithme LOP. En effet, les MP initiaux sont créés via un processus lourd qu'il n'est pas possible de modifier à court terme. Par ailleurs, l'algorithme LOP a nécessité des investissements importants et les contraintes des clôtures IFRS 17 font qu'il n'est pas envisageable de le remplacer à court terme.

Après avoir rappelé les exigences introduites par IFRS 17 et les évolutions de l'outil de modélisation, nous procéderons à une description des données du portefeuille de prévoyance individuelle concernée. Une analyse du modèle d'agrégation nous permettra par la suite d'identifier les leviers d'optimisation. Ce qui nous permettra de proposer un calibrage de ce modèle d'agrégation, calibrage basé sur le choix des variables pour le calcul de la fonction cible et d'un pré clustering sur les composantes principales pour les variables de regroupement. Cette agrégation permettra d'obtenir des résultats de bien meilleure qualité que celle mise en œuvre lors des clôtures précédentes.

1. CONTEXTE DE L'ETUDE ET PROBLEMATIQUE

La norme IFRS 17 introduit un changement profond chez les assureurs. Plus que de nouveaux concepts, elle introduit en effet une nouvelle manière de reconnaître le résultat d'assurance à une maille plus fine. Cette exigence implique la revue des processus et la modification des outils de modélisation. Dans ce chapitre nous présenterons la norme IFRS 17 en insistant sur la question de la granularité des données. Nous présenterons ensuite l'évolution de l'outil de modélisation qui s'est imposée, ce qui nous permettra de présenter la problématique de notre étude.

1.1 L'introduction de la norme IFRS 17

1.1.1 Présentation et principaux enjeux

Le projet de norme sur les contrats d'assurance a débuté à la fin des années 90. Sa date d'entrée en vigueur a plusieurs fois été reportée. Le schéma ci-après montre les différentes étapes qu'a connues cette norme depuis sa publication en mai 2017.

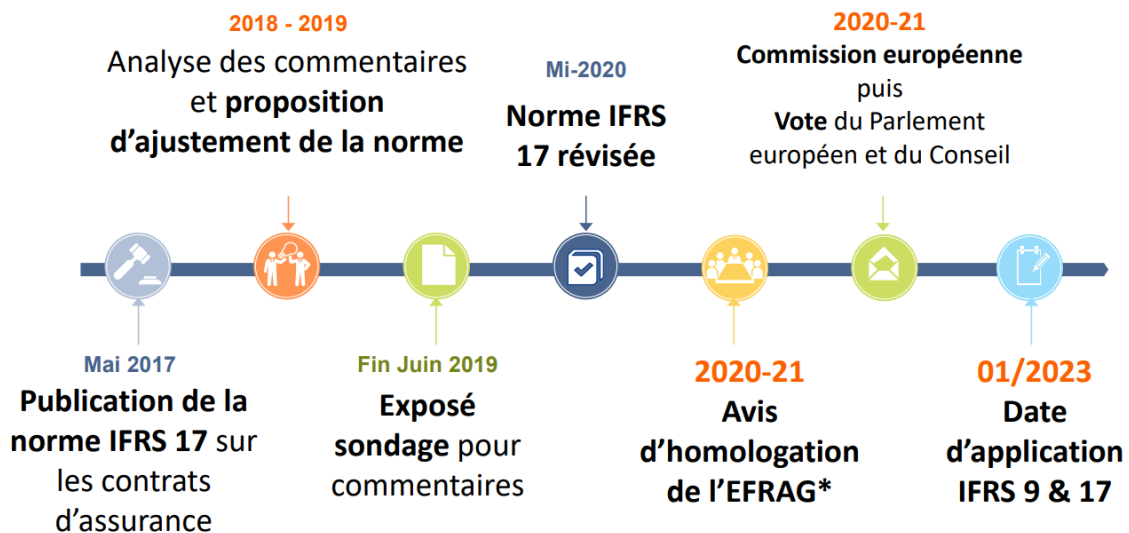


Figure 1 : les différentes étapes de la norme IFRS 17 depuis sa publication

1.1.1.1 Présentation générale de la norme

L'objectif de la norme IFRS 17 est d'améliorer la comparabilité des états financiers entre sociétés en imposant un modèle qui reflète davantage la réalité économique des contrats

d'assurance. IFRS 17 s'inspire des approches de valorisation des passifs d'assurance proposées par S2, introduit la notion de Marge de services contractuels (Contractual Service Margin – CSM) et requiert de nouvelles exigences en termes de maille de calcul (le groupe de contrats).

Lors de la comptabilisation initiale, l'organisme doit évaluer chaque groupe de contrats d'assurance comme la somme des deux éléments suivants :

- Les flux de trésorerie d'exécution (Fullfilment cash flows), constitués des éléments suivants :
 - o Les estimations des flux de trésorerie futurs, assimilables à une provision Best Estimate des engagements (BEL) de l'organisme à l'égard des assurés du groupe de contrats ;
 - o Un ajustement destiné à refléter la valeur temps de l'argent et les risques financiers liés aux flux de trésorerie futurs ;
 - o Un ajustement au titre du risque non financier (Risk Adjustment – RA)
- La marge sur services contractuels (Contractual Service Margin - CSM), représente les profits futurs non acquis que l'organisme s'attend à reconnaître au fur et à mesure qu'il fournira les services associés au contrat.

Le schéma suivant compare les postes du passif du bilan IFRS 4 au nouveau bilan IFRS 17.

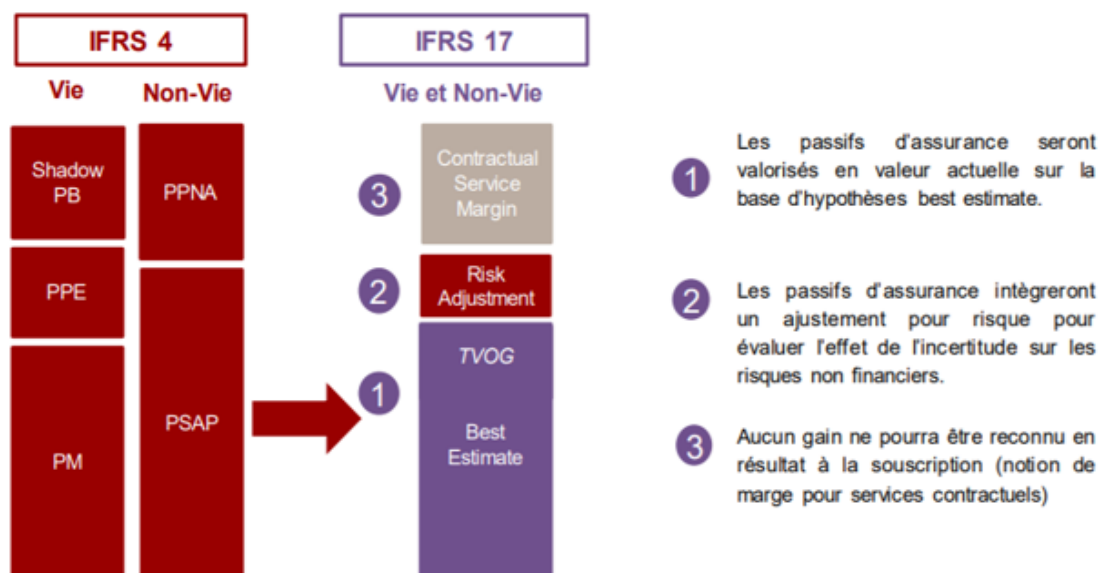


Figure 2 : Comparaison bilans IFRS 4 vs IFRS 17

Nous présentons par la suite chacun de ces nouveaux agrégats.

- **BEL**

Le *Best Estimate Liabilities* (BEL) correspond à l'estimation explicite, objective et pondérée par les probabilités (c'est-à-dire l'espérance mathématique) de l'écart de la valeur actualisée des sorties de trésorerie futures sur la valeur actualisée des entrées de trésorerie futures qui découleront de l'exécution des contrats d'assurance par l'organisme.

Globalement, les principes sont assez similaires à ceux retenus dans l'évaluation du *Best Estimate* des provisions sous Solvabilité 2. Les estimations des flux de trésorerie doivent :

- intégrer avec objectivité l'ensemble des informations raisonnables et justifiables qu'il est possible d'obtenir sans coût ou effort excessif sur le montant, l'échéancier et l'incertitude des flux de trésorerie futurs ;
- tenir compte de l'incertitude des flux ; pour ce faire, l'organisme doit estimer l'espérance mathématique (c'est-à-dire la moyenne pondérée selon les probabilités) de l'éventail complet des résultats possibles ;
- refléter le point de vue de l'organisme, pour autant que les estimations des variables de marché pertinentes cadrent avec les prix de marché observables pour ces variables ;
- être à jour : les estimations doivent refléter les conditions existant à la date d'évaluation, notamment les hypothèses à cette date quant à l'avenir ;
- être explicites : l'organisme doit établir l'ajustement au titre du risque non financier séparément des autres estimations. L'organisme doit aussi estimer les flux de trésorerie séparément de l'ajustement au titre de la valeur temps de l'argent et du risque financier

- **Ajustement pour risque non financier**

L'ajustement au titre du risque non financier ou *Risk Adjustment* (RA) doit refléter l'indemnité qu'exige l'organisme pour la prise en charge des risques non financiers associés aux flux de trésorerie impactant le montant et l'échéancier des flux de trésorerie. Il doit également refléter :

- le bénéfice de diversification au niveau de l'organisme d'assurance,
- le degré d'aversion au risque de l'organisme.

La norme n'impose pas de méthode particulière de détermination du RA, mais précise que le montant de l'ajustement doit présenter les caractéristiques suivantes (IFRS 17.B91) :

- être plus élevé si les risques sont moins fréquents mais graves que s'ils sont fréquents mais peu graves, c'est-à-dire que l'ajustement sera plus élevé pour des risques liés aux sinistres extrêmes que pour des sinistres attritionnels ;
- être plus élevé si la durée des contrats est plus longue à risque similaire ;
- être plus élevé si la distribution de probabilité des risques est large que si elle est étroite ;
- être plus élevé si l'estimation et la tendance présentent de nombreuses inconnues ;
- être moins élevé si les récents résultats techniques réduisent l'incertitude autour du montant et de l'échéancier des flux de trésorerie, et vice-versa.

Trois approches sont pratiquées pour déterminer le RA³ :

- l'approche par le coût du capital : le RA est la somme des SCRs (*Solvency Capital Requirement*) actualisés pendant toute la période de couverture (pour les risques non financiers) actualisés en multipliant avec un taux du coût du capital ;
- l'approche par les chocs : $RA^k = BEL^k - BEL$, ou RA^k *Risk Adjustment* pour un risque non financier k et BEL^k , le BEL calculé après application d'un choc 'ajusté' sur le facteur de risque k , selon la méthodologie que la formule standard S2 ;
- l'approche stochastique : le calcul du RA s'appuie sur un Générateur de Scénarios Économiques (GSE). L'entité projette les flux de trésorerie futurs en diffusant des scénarios de risques non financiers afin d'obtenir une distribution des flux futurs actualisés. La distribution ici est engendrée par l'aléa sur les risques non-financiers, les risques financiers étant 'fixés'. Grâce à cette distribution, l'entité déduit la valeur du RA en calculant une mesure de risque telle que la *Value at Risk* ou la *Tail Value at Risk*.

La norme ne précise pas l'approche à retenir mais demande à l'entité de communiquer le seuil de confiance associé au calcul du RA.

Enfin, le RA doit refléter le niveau de diversification au sein de l'entité dans l'évaluation de l'incertitude sur les cash-flows. Il peut donc intégrer les effets de diversification pouvant exister

³ Nexialog Consulting [2021], *IFRS 17 Risk Adjustment, Une étude comparative sur le périmètre des contrats d'épargne en euros*, Paris, France.

au sein d'un groupe d'assurance dans la mesure où des mécanismes de solidarité financière existent. Si de tels mécanismes n'existent pas au sein du groupe, la diversification devra être considérée au niveau des entités du groupe (au maximum) (Transition Ressource Group : publication du 05/2018 – AP02)

- **La marge future est externalisée au travers de la CSM**

La CSM, *Contractual Service Margin*, est un nouveau concept introduit par IFRS 17 pour identifier la profitabilité des groupes de contrats et la reconnaître progressivement au cours du temps. La marge de service contractuelle a pour but d'éliminer tout gain à la souscription et représente le profit non acquis que l'entité s'attend à reconnaître en fonction de l'écoulement du service. Economiquement, la CSM correspond aux profits attendus non reconnus du contrat avant impôts sur l'horizon de couverture. Elle est inscrite au passif du bilan si elle est positive (sinon, une perte est comptabilisée en compte de résultat). Lors de la première comptabilisation d'un contrat, la CSM est égale au montant opposé du flux issu de la somme des éléments suivants :

- les flux de cash flows (BE + RA) ;
- les frais d'acquisition déjà payés (frais de pré-couverture) ;
- tout autre flux issu des contrats du groupe à cette date de première comptabilisation.

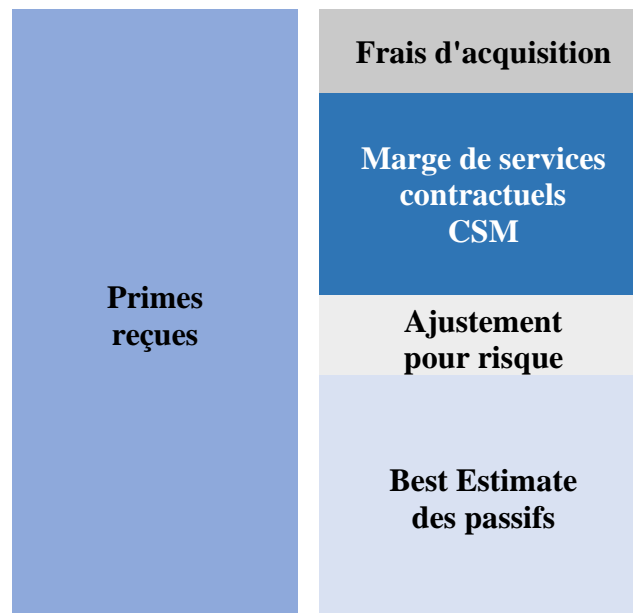


Figure 3 : Détermination de la CSM

1.1.1.2 Exigences en termes de maille de calcul⁴

- Généralités

La CSM doit être évaluée à une granularité particulière définie par la norme : on parle d'unité de mesure ou de groupe de contrats. L'objectif est de limiter les effets de compensation entre les contrats. La norme indique que les contrats d'assurance peuvent être agrégés selon les trois critères ci-dessous :

- portefeuille de contrats ;
- rentabilité ;
- cohorte annuelle (génération).

Le premier critère vise à prohiber la compensation des gains et pertes entre des types de produits différents. Le second critère permet de limiter les effets de mutualisation au sein des mêmes types de produits en séparant les contrats a priori rentables des contrats a priori onéreux. Enfin la notion de génération interdit les effets de compensation entre des contrats émis à plus d'un an d'intervalle. La classification des contrats selon les trois critères ci-dessus conduit à la définition de groupes de contrats (ou unités de mesure). Cette classification est effectuée a priori lors de la souscription des contrats et n'est pas revue par la suite.

- Notion de portefeuille

Un portefeuille au sens d'IFRS 17 regroupe des contrats sujets à des risques similaires et gérés ensemble. La notion de risques similaires est définie comme le fait que les cash flows futurs soient similaires dans leur montant et « *timing* » entre deux contrats si les hypothèses sous-jacentes sont proches.

La notion de gestion unique des contrats intègre notamment les aspects liés à la tarification : par exemple des contrats étant commercialisés de manière différenciée (réseau de distribution différent) mais dont la gestion et la tarification sont régies par des règles communes (exemple : existence de guidelines, etc.) pourront être regroupés.

⁴ Cette partie s'inspire de la présentation de la norme effectuée par Renaud BAILLY et Nicolas GEMIN [2018], *IFRS 17 : interprétation de la norme, premiers résultats et leviers de pilotage pour un portefeuille dommages*, Mémoire d'actuariat, CEA

En pratique les contrats faisant partie des mêmes lignes de produits pourront être regroupés s'ils sont gérés de manière unique et sont sujets à des risques de même nature.

- Séparation des contrats selon leur profitabilité attendue

Au sein d'un portefeuille de contrats, la norme impose qu'une entité répartisse les contrats en trois classes :

- Un groupe de contrats onéreux à la souscription (contrats dont la CSM est nulle et qui génèrent une perte immédiate à la première comptabilisation à hauteur du déficit à comptabiliser directement au compte de résultat) ;
- Un groupe de contrats profitables à l'émission ayant peu de probabilité de devenir onéreux ;
- Un groupe avec le reste des contrats du portefeuille.

Ces regroupements doivent être effectués sur la base des éléments à la disposition de l'entité lors de la reconnaissance du contrat. La norme décrit deux manières possibles d'effectuer ces regroupements :

- Si l'entité possède des informations « raisonnables et justifiables » lui permettant de conclure qu'un ensemble de contrats appartient au même groupe, elle peut évaluer à quel groupe cet ensemble appartient au global (et non contrat par contrat) ;
- Dans le cas contraire elle devra évaluer contrat par contrat à quel groupe de profitabilité appartient chaque contrat individuel.

Pour les contrats qui ne sont pas onéreux à la souscription, l'entité devra les répartir entre les classes « profitable sans risque de devenir onéreux » ou « profitable avec risque de devenir onéreux ». Pour ce faire, l'entité devra évaluer qu'un contrat n'a pas de risque de devenir onéreux :

- en se basant sur la probabilité des changements d'hypothèses qui, s'ils arrivaient, conduiraient à rendre le contrat onéreux ;
- en se basant sur les informations disponibles au sein des systèmes d'information de l'entité. Si certains contrats d'un même portefeuille doivent être répartis au sein de groupes différents, uniquement en raison de contraintes légales ou réglementaires sur la tarification, la norme autorise à ce qu'ils soient regroupés. C'est le cas par exemple pour l'interdiction de différencier les tarifs entre homme et femme : si cette contrainte

entraîne, par exemple, le fait que les contrats souscrits auprès d'hommes soient déficitaires ils pourront être regroupés avec les contrats souscrits auprès de femmes (la profitabilité de l'ensemble de la population sera alors appréciée).

En pratique, la profitabilité devra être évaluée en utilisant les éléments liés à la tarification des produits, la réalisation des exercices budgétaires ainsi que les reportings suivis pour piloter la compagnie.

- Séparation des cohortes annuelles

La norme interdit de regrouper les contrats émis à plus d'un an d'intervalle. Cette interdiction doit s'entendre en prenant en compte la frontière des contrats (exemple : pour un contrat annuel tacitement reconductible, la date de souscription ne correspondra pas à la souscription initiale mais à la date du dernier renouvellement).

Cette restriction a été introduite par l'IASB, l'*International Accounting Standards Board*, car elle craignait qu'en l'absence de cette contrainte les assureurs définissent des groupes de contrats éternellement ouverts. Ceci aurait pu conduire à une perte d'information sur l'évolution de la profitabilité au cours du temps.

En pratique la mise en œuvre de cette exigence est complexe pour les entreprises. C'est pour cela que le régulateur européen dans l'amendement N°2021/2036 du 19 novembre 2021, rend cette exigence optionnelle. Les compagnies d'assurance qui exercent cette option doivent cependant l'indiquer dans leurs états financiers.

Seules les deux premières exigences, la notion du portefeuille IFRS 17 et la notion de profitabilité, sont appliquées dans le portefeuille que nous étudions. La prise en compte de la troisième exigence est en cours de discussion à la date de rédaction de ce mémoire. Si elle était prise en compte, la question de l'agrégation des données serait encore plus nécessaire.

Les exigences en termes de maille de calcul étant présentées, nous allons terminer cette partie en présentant les 3 modèles d'évaluation possibles en IFRS 17.

1.1.1.3 Les 3 modèles d'évaluation possibles

La norme IFRS 17 définit trois modalités de comptabilisation des contrats d'assurance en fonction de leur nature, on parle de modèle comptable.

- **Modèle général**

Le modèle général ou BBA (Building Block Approach) s'applique par défaut à tous les contrats en dehors des contrats dits « participatifs directs ». Ce modèle s'applique donc aux types de contrats suivants :

- contrats non participatifs : contrats dont les éléments sous-jacents appartiennent à l'assureur et celui-ci ne fait participer les assurés ni au résultat financier ni au résultat technique (contrats dommages, contrats de prévoyance type temporaire décès, etc.) ;
- contrats participatifs indirects : contrats dont les éléments sous-jacents appartiennent à l'assureur qui dispose d'une entière discrétion sur la participation accordée aux assurés (exemple : certains contrats de retraite, contrats type vie entière, etc.).

- **Modèle simplifié (PAA)**

Le modèle simplifié PAA (Premium Allocation Approach) est un modèle comptable optionnel applicable à tous les contrats sous certaines conditions. La condition d'éligibilité à la PAA est composée de deux critères non-cumulatifs lors de l'émission des contrats :

- les résultats obtenus par le biais de l'approche simplifiée ne diffèrent pas de façon matérielle de ceux qui seraient obtenus dans le modèle général ;
- ou, la période de couverture des contrats d'assurance à la date de première comptabilisation (au regard de la frontière des contrats) est inférieure ou égale à 1 an.

Le premier critère n'est pas respecté si, à la souscription du groupe de contrats, une entité s'attend à une variabilité significative des engagements avant la survenance des sinistres. Cette variabilité concerne par exemple :

- l'ampleur des flux de trésorerie futurs liés aux dérivés incorporés ;
- la durée de couverture du groupe de contrats. De plus, l'entité devra interpréter le terme « de façon matérielle ».

Les éléments pouvant conduire à une différence entre PAA et modèle général peuvent être les suivants :

- modalités de reconnaissance de la CSM et des cash-flows prévisionnels différents du passage du temps (saisonnalité) ;
- effets liés à l'actualisation.

Le second critère signifie que tous les contrats ayant une période de couverture d'un an ou moins sont éligibles à la méthode PAA

- Modèle VFA

Pour les contrats dits « participatifs directs », l'entité doit appliquer un modèle comptable spécifique dit *Variable Fee Approach* (VFA).

Ce modèle consiste à rapprocher le contrat d'assurance d'un élément financier ayant une composante d'investissement et à considérer que tout prélèvement sur cette composante constitue des frais variables pour rémunérer les services fournis. La principale différence entre le modèle VFA et le modèle général est que la CSM est impactée des variations du contexte économique dans le modèle VFA contrairement au modèle général. Ainsi, un contrat participatif comptabilisé en modèle général verra l'intégralité des variations de marges futures liées à des changements économiques impacter directement le compte de résultat. Un contrat participatif comptabilisé en VFA verra lui ces variations impacter la CSM, l'effet ne sera ainsi impacté en résultat qu'au travers de l'amortissement de celle-ci.

Les critères d'éligibilité à la VFA sont les suivants :

- (a) les termes contractuels spécifient que l'assuré bénéficie d'une part d'éléments sous-jacents clairement identifiés ;
- (b) l'assureur s'attend à payer à l'assuré un montant égal à une part substantielle du rendement en *fair value* des éléments sous-jacents ;
- (c) l'assureur s'attend à ce que les variations de l'actif sous-jacent soient répercutées en proportion substantielle à l'assuré.

Les caractéristiques de participation des contrats doivent conduire à ce que les obligations de l'assureur envers ses assurés correspondent au net entre :

- le paiement à l'assuré d'un montant égal à la juste valeur des sous-jacents et,
- les chargements prélevés en vertu des services fournis par le contrat d'assurance incluant :
 - une part de la juste valeur des actifs.
 - des flux de trésorerie qui n'évoluent pas selon les rendements des sous-jacents (frais, etc.).

La classification en VFA se fait à l'émission des contrats et n'est pas réévaluée. L'effet des taux garantis est ainsi à considérer à l'aune des conditions financières observées à la souscription des contrats. De plus, la « part » du pool mentionnée dans le (a) ne se limite pas à la seule discrétion de l'assureur mais impose qu'il existe des contraintes juridiques, contractuelles ou réglementaires de verser cette part. Enfin, les « rendements » et « variations » mentionnés dans les conditions (b) et (c) s'entendent sur la durée des contrats et dans une logique de valeur actuelle probable moyenne sur plusieurs scénarios financiers.

Les données de notre portefeuille sont valorisées selon le modèle VFA.

1.2 L'évolution du modèle ALM.

Après avoir présenté les limites du modèle ALM actuel (Flexing), nous présenterons le nouveau modèle *Full* ALM.

1.2.1 Les limites du modèle *flexing* actuel (ALIM)

1.2.1.1 Modèle déterministe (MoDet)

Le modèle ALM actuel utilise en entrée des flux générés par un modèle élémentaire de passif. Il s'agit d'un modèle actuariel déterministe à pas mensuel utilisé par exemple pour effectuer des prévisions financières, des études actif/passif, des calculs d'*Embedded Value*.

Sa structure est la suivante :

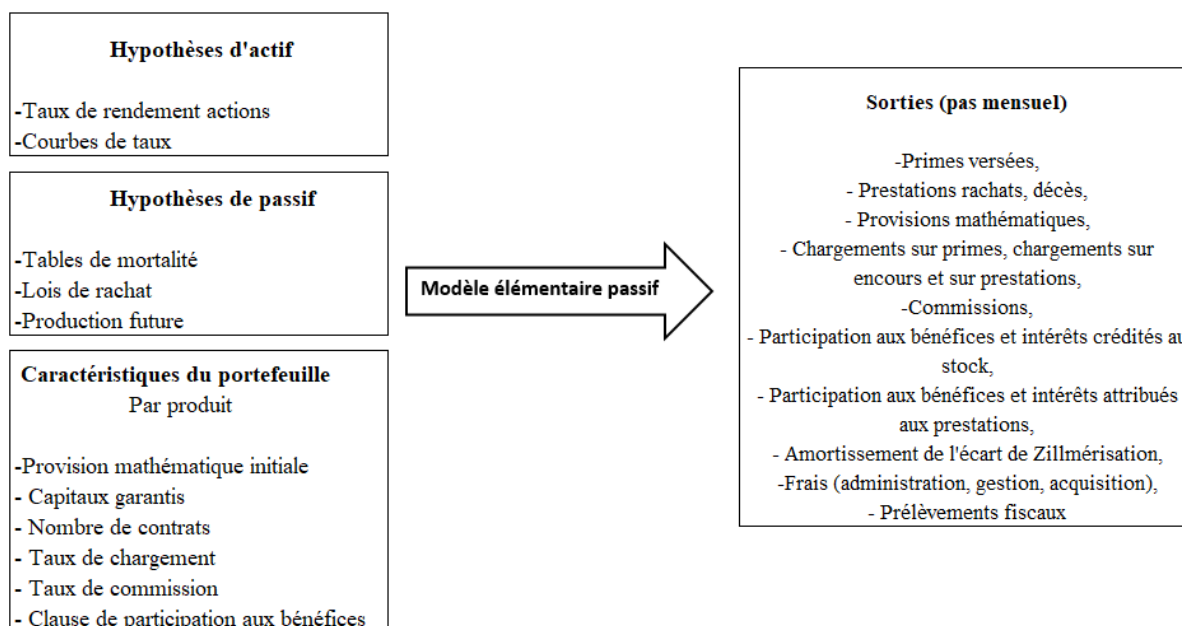


Figure 4 : Structure d'un modèle déterministe

Les provisions mathématiques de clôture sont calculées par récurrence à partir des flux de la période. Les prestations sont calculées à partir des lois de rachat et des tables de mortalité paramétrées dans le modèle.

Il ne s'agit donc pas d'un modèle actif/passif intégré, puisqu'il n'y a pas de modélisation dynamique des actifs et pas de modification du comportement des assurés en fonction des conditions de marché.

En revanche, son atout est qu'il permet d'avoir une bonne connaissance des produits et le choix d'un pas mensuel permet un niveau très fin de modélisation. Par exemple, cela permet de distinguer l'attribution de la participation aux bénéfices selon la clause exacte du contrat : au 1^{er} janvier, au 1^{er} juillet ou à date anniversaire.

1.2.1.2 La nécessité d'un modèle *flexing*

Il n'est pas possible d'utiliser directement le modèle déterministe de passif dans le cadre d'un modèle ALM. D'une part à cause du temps de run. En effet, si le temps de simulation du modèle passif élémentaire est de 30 secondes, le temps de run pour une simulation de 1000 scénarios sera supérieur à 8 heures ! D'autre part, l'utilisation du modèle de passif impliquerait de faire de nombreuses modifications dans le code pour l'alimentation des différents taux et rendements et pour la prise en compte de l'impact des conditions de marché sur les rachats, par exemple.

Une des solutions possibles à ces problèmes est d'utiliser la technique du *flexing*. Au lieu de recalculer à chaque période les prestations, provisions et autres données du passif, on déforme des flux pré-simulés, issus du modèle de passif détaillé, en fonction des conditions de marché fournies par le modèle d'actif. Concrètement :

- On récupère les flux de passif dans une hypothèse de projection au taux technique (sans option de PB) du produit que l'on souhaite modéliser, à partir des résultats d'un modèle élémentaire développé par les actuaires de la société étudiée ;
- Le modèle bilantiel détermine en fonction des revenus de l'année et des conditions de marché le montant de PB incorporé aux contrats et déclenche les éventuels rachats conjoncturels ;
- Ces flux et ces montants bilantiels permettent de calculer les ratios d'ajustement ;
- On en déduit les flux de passifs déformés en multipliant les flux de passifs originaux par le ratio correspondant.
- Et on répète l'opération à chaque pas de temps.

Le schéma ci-après présente le processus de *flexing* actuellement utilisé dans le cadre de la modélisation ALM.

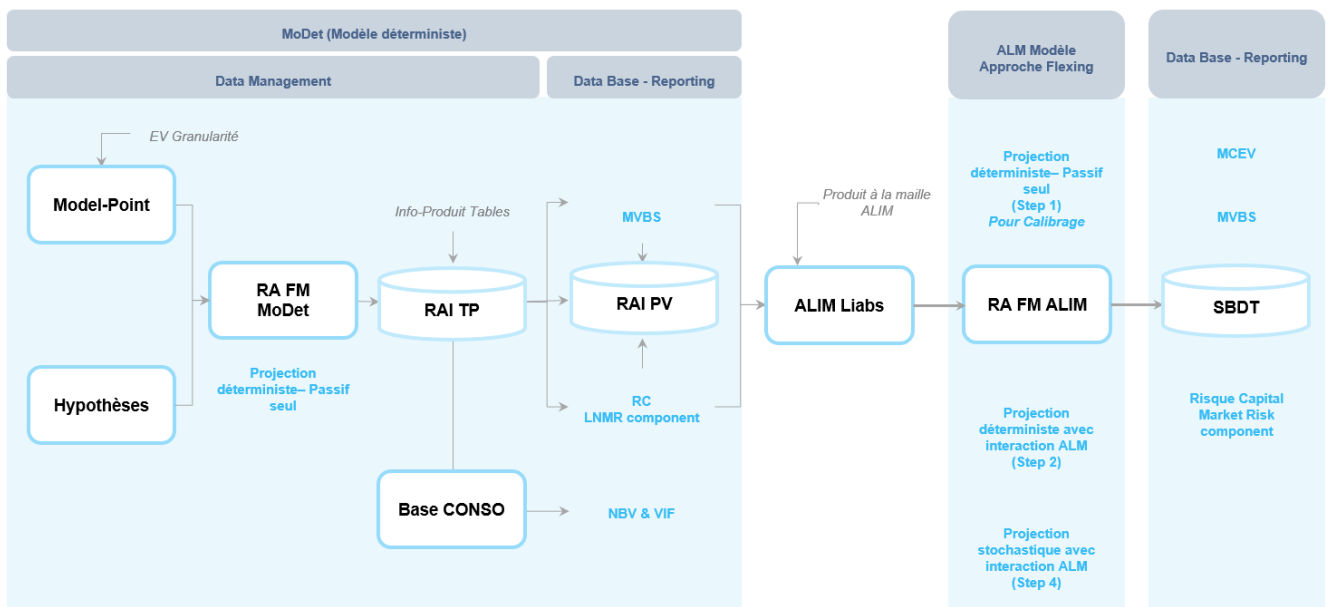


Figure 5 : Processus *flexing*

Les MP et les hypothèses permettent d'effectuer les projections déterministes en mode passif seul, c'est-à-dire sans interaction avec l'actif. Ces résultats alimentent RAI TP (Risk Analysis Infrastructure –Technical Provision) une base de données qui retraite les flux déterministes

projetés pour calculer la valeur des affaires nouvelles (NBV) et la valeur des affaires en portefeuille (VIF). Après divers traitements de données (agrégation, annualisation, chargements de tables, etc) les fichiers de passifs, appelés « produits ALIM », nécessaires au fonctionnement du modèle ALIM sont créés via une autre base de données (Risk Analysis Infrastructure- Passif vie). Ces fichiers avec les MP actifs, les scénarios et autres inputs permettent de lancer le modèle ALIM en step 1 (calibrage), en step 2 (projection déterministe avec interaction ALM) et en step 4 (projection stochastique avec interaction ALM). Les résultats de la modélisation ALM permettent d'effectuer les reportings MCEV (Market Consistency Embedded Value), MVBS (Market Value Balance Sheet), risque capital.

1.2.1.3 Modélisation stochastique

Le modèle stochastique, ALIM (Assets and Liabilities Interaction Management), permet de modéliser les interactions entre l'actif et le passif. Il est utilisé pour reproduire le fonctionnement et simuler l'évolution du bilan d'une compagnie d'assurance selon différents scénarios économiques sur un horizon de temps défini, en projetant l'ensemble de ses cash-flows. Cet outil permet l'intégration de la stratégie financière (revalorisation future des contrats) ; les aspects commerciaux et les engagements contractuels en prenant en compte le comportement dynamique des assurés, notamment les options et garanties.

- Présentation des modules du modèle ALM

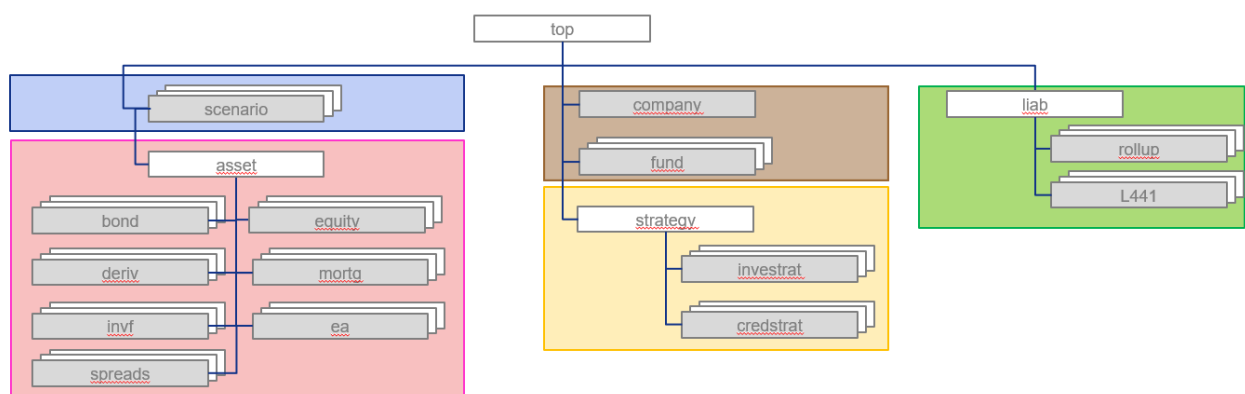


Figure 6 : Modules du modèle ALIM

La figure ci-dessus présente l'architecture du modèle actif-passif, qui est divisé en plusieurs sous-modules. Chacun des sous-modules regroupe les principaux inputs nécessaires aux calculs des engagements d'assurance-vie :

- Le module « **Scénario** » permet de prendre en compte différents scénarios économiques (prix des obligations zéro coupon pour différentes maturités et pour différentes monnaies, rendement des différents indices actions contenus en portefeuille, rendement de l'immobilier, valeur des différents taux de change, etc). Chaque scénario est composé de plusieurs trajectoires économiques identiques au pas de temps 0, et qui, en évoluant de manière aléatoire, vont former une distribution obéissant à des probabilités risque neutre. Il permet de projeter l'actif et le passif selon ces différentes hypothèses économiques.
- Le module « **Asset** » est quant à lui composé des différents portefeuilles d'actifs en stock à la date d'évaluation, qui sont les données d'entrée. Les actifs sont agrégés par type, par caractéristiques homogènes et par fonds. Par exemple, le sous-module nommé « bond » sur ALIM, va se décomposer en Fixed income, Loans, Floating Rate Notes, Inflation Linked Bonds et Bond Forward, qui sont tous des instruments ayant des caractéristiques communes. Ce module permet la modélisation des cash flows d'actifs permettant d'obtenir les revenus courants pour le calcul des produits financiers, la valorisation en valeur de marché des actifs pour la partie « Actif » du bilan économique Solvabilité II.
- Le module « **Company** » : ce module permet l'agrégation des données pour les calculs à effectuer au niveau fonds et/ou compagnie
- Le module « **Strategy** » permet de modéliser la stratégie d'investissement et la politique de participation aux bénéfices (pour les contrats Euro). La stratégie d'investissement va permettre de définir l'allocation cible du portefeuille d'actifs en fonction de l'environnement économique. Cette stratégie a un impact sur les actifs du bilan. D'après l'article L331-3 du code des assurances, « les entreprises d'assurance sur la vie doivent faire participer les assurés aux bénéfices techniques et financiers qu'elles réalisent ». Ainsi, les assureurs sont légalement tenus de redistribuer aux souscripteurs de contrats Euro, ce qui a été garanti, mais également une partie de leurs bénéfices. La Crediting Strategy a donc pour principal objectif le calcul d'un taux de participation aux bénéfices cible, qui est obtenu en prenant en compte (pour des raisons commerciales) un taux de marché et le taux crédité en $t - 1$. Cette stratégie interagit avec le passif du bilan.

- Le module « **Liability** » permet de modéliser les flux de passifs issus des modèles déterministes (qui sont agrégés en fonction des caractéristiques des produits telles que le taux de participation aux bénéfices par exemple) en les déformant dans ALIM en fonction de l'environnement économique. Les modèles déterministes vont modéliser les flux de chaque produit d'assurance selon un scénario dit « central » (ayant seulement une seule trajectoire) qui correspond à la situation du marché à la date d'évaluation, et qui représente l'évolution de l'ensemble des paramètres du marché. En sortie, le modèle retourne pour chaque produit, tous ses cash-flows sur les soixante prochaines années. Concrètement, la probabilité qu'un tel scénario se produise est nulle, c'est pour cela que des modèles stochastiques sont utilisés permettant de capter la « valeur temps » de nos indicateurs, mais également d'introduire des lois comportementales de l'assuré qui sont fortement dépendantes de l'environnement économique. Étant donné que les risques du passif dépendent fortement de l'assuré, l'outil ALIM est intéressant car il permet, dans son calcul des flux de passif, la prise en compte des rachats dynamiques, des liens entre produits (arbitrages, transferts de rentes), et de la revalorisation des contrats Euro (taux de participation aux bénéfices qui correspond au taux crédité par la Crediting Strategy). La projection des cash-flows de passif évolue donc selon le contexte financier et le comportement des clients.

1.2.2 La nécessité de passer à un modèle *Full* ALM (LSIM)

1.2.2.1 Objectifs

Dans le cadre du déploiement d'un modèle *full* ALM pour les travaux IFRS 17, le modèle LSIM, *Local Stochastic Integrated Model*, a été développé pour répliquer en un modèle unique les calculs effectués sur les modèles existants :

- Réplication de la modélisation passif MoDet jugée plus appropriée afin de pallier les limites observées dans le modèle ALIM.
- Réplication des mécanismes actif/passif et de la fonctionnalité stochastique dans ALIM.
- En combinant les avantages de la modélisation MoDet et ALIM, le modèle LSIM pourra projeter directement, plus finement (via les *model points*) et de manière plus adéquate le portefeuille tout en prenant en compte les interactions ALM de manière stochastique.

Ce modèle a un pas annuel.

- Architecture du modèle

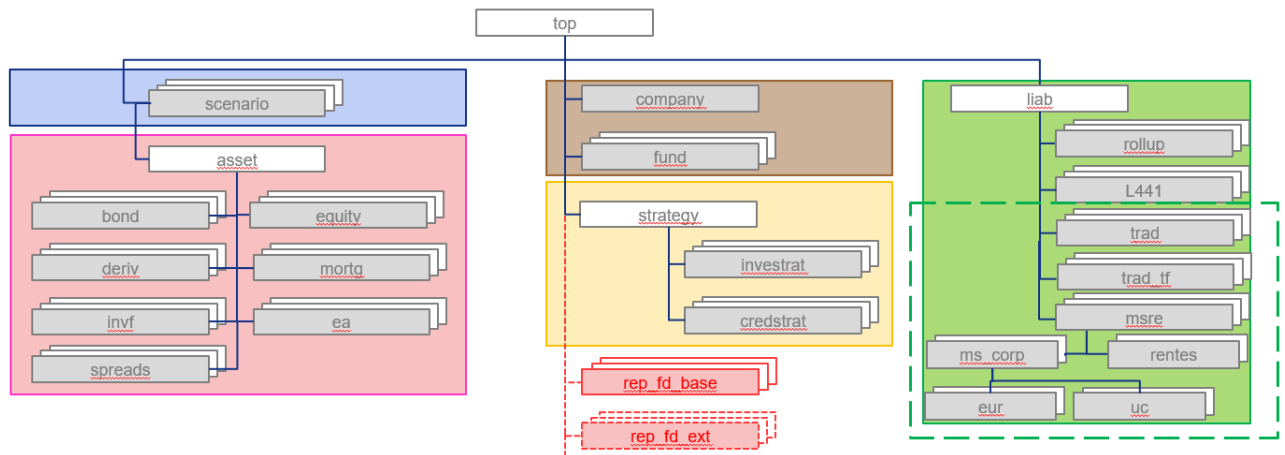
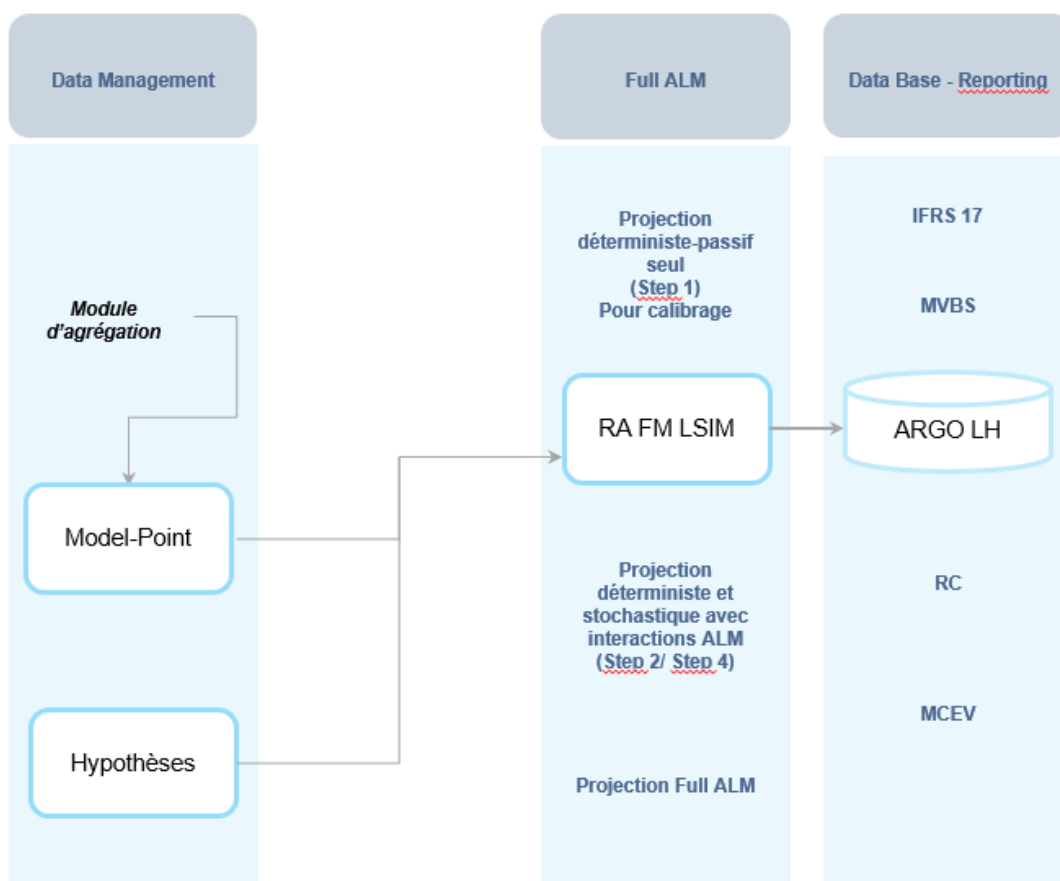


Figure 7 : Architecture du modèle ALM

Le modèle LSIM est composé de 7 sous-modèles de passif supplémentaires (par rapport au modèle ALIM) reflétant l'intégration d'une modélisation type modèle déterministe :

- Sous-modèle prévoyance (Trad), pour la modélisation des produits de prévoyance individuelle longue.
- Sous-modèle termes fixes (Trad_TF) pour la modélisation des produits dit « Traditionnels Termes Fixes ».
- Sous-modèle d'agrégation de l'épargne et des rentes (MSRE). Ce sous-modèle comporte :
 - Sous-modèle rentes en phase de restitution
 - Sous-modèle d'agrégation de l'épargne et des rentes en phase de constitution
 - Sous-modèle épargne euro (et rentes en phase de constitution)
 - Sous-modèle épargne UC (et rentes en phase de constitution)

Le schéma ci-après présente le processus de *full* ALM utilisé dans le cadre de la modélisation LSIM.

Figure 8 : Processus *Full ALM*

Comparé au processus *flexing*, le processus *full ALM* est simplifié et la modélisation s'effectue sur la base des *model points* au lieu des poches de passif (liabilities) en *flexing*.

1.3 Problématique de l'étude

Nous avons vu précédemment que la norme IFRS 17 exige un niveau de calcul, d'analyse et de reporting plus fin des contrats d'assurance. Le passage du modèle *flexing* au modèle *Full ALM* permet de répondre à ces exigences. Le processus actuariel est raccourci (moins d'étapes) et les calculs sont effectués sur la base d'une maille plus fine, *model point* (MP) par *model point* (~ 120 000 MP) au lieu de la maille ALIM (~ 500 produits).

Cependant, avec une maille si fine, les temps de run d'un modèle ALM sont très importants. Il faut donc agréger les MP sans trop perdre d'information avant de lancer le modèle ALM. C'est-

à-dire agréger avec une erreur relative acceptable sur les BEL (Best Estimate Liabilities) et la PVFP (Present Value of Future Profit).

Un algorithme (LOP) a été mis en place pour effectuer cette agrégation. Sur le périmètre prévoyance individuelle, durant les premières clôtures à blanc des comptes IFRS 17, la qualité d'agrégation n'était pas satisfaisante : avec 18 525 *model points* (MP) initiaux, l'algorithme permettait d'obtenir 4 056 MP agrégés, soit une réduction de 78%. Le BEL initial était de 818 millions €, le BEL après agrégation de 787 millions €, soit une erreur d'estimation de 3,78%.

L'objectif de notre étude est d'améliorer la qualité d'agrégation sur ce périmètre sans modifier les processus de création de MP et sans modifier l'algorithme LOP. En effet, les MP initiaux sont créés via un processus lourd qu'il n'est pas possible de modifier à court terme. Par ailleurs, l'algorithme LOP a nécessité des investissements importants et les contraintes des clôtures IFRS 17 font qu'il n'est pas envisageable de le remplacer à court terme.

Pour atteindre notre objectif, nous allons premièrement présenter l'algorithme LOP qui repose sur la technique des moindres carrés non négatifs (NNLS). Ceci nous permettra d'identifier ces leviers d'optimisation : les variables de calcul de l'algorithme (matrice LOP) et les variables de regroupement des inputs de l'algorithme.

Pour les variables de calcul nous calibrerons les variables LOP qui permettent de calculer les valeurs cibles au travers des corrélations sur des indicateurs (BEL, PVFP, BEL et PVFP).

Pour les variables de regroupement nous utiliserons deux approches pour les calibrer. Une première approche consistera à appliquer un algorithme de classification non supervisée, le K-Means, basé sur les principaux cash-flows de passif déterministes avant agrégation. Dans une seconde approche, on appliquera le même algorithme de classification, mais cette fois-ci sur les MP. Dans ce dernier cas, compte tenu de la dimension des MP (13 variables quantitatives), il faudra au préalable effectuer une analyse en composantes principales (ACP). Un théorème énoncé par Ding & He (2004)⁵, nous permettra d'effectuer la combinaison entre le K-Means et l'ACP et de trouver le bon nombre de composantes principales à utiliser. Une comparaison des résultats selon chaque approche et entre les variantes de chaque approche nous permettra de proposer un calibrage du modèle.

⁵ Ding C & He X [2004], *K-means Clustering via Principal Component Analysis*, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

2. PRESENTATION DES DONNEES

Ce chapitre décrit les données utilisées dans le cadre de ce mémoire. Le portefeuille étant constitué des produits de prévoyance individuelle, nous commencerons dans une première partie par définir la prévoyance individuelle et les garanties de notre portefeuille. Dans une seconde partie nous présenterons les informations du portefeuille grâce aux statistiques descriptives.

2.1 Prévoyance individuelle

2.1.1 Généralités

En France, les salariés titulaires d'un contrat de travail sont couverts par la Sécurité Sociale pour les risques maladie, maternité, invalidité, vieillesse et décès. La prévoyance est une protection sociale complémentaire à celle de la Sécurité Sociale. Plus précisément, la prévoyance recouvre tout ce qui concerne la protection des personnes, à savoir :

- Le décès,
- L'incapacité,
- L'invalidité,
- La maternité,
- L'hospitalisation,
- La dépendance.

La souscription des garanties peut être effectuée par une personne morale, par exemple une entreprise pour ses salariés : c'est la prévoyance collective. Elle peut aussi être effectuée directement par une personne physique : c'est la prévoyance individuelle. C'est la prévoyance individuelle qui nous intéresse dans le cadre de ce mémoire.

2.1.2 Types de garanties

La prévoyance individuelle recouvre les risques suivants : décès, incapacité et invalidité.

- Décès

Les garanties proposées en cas de décès peuvent être le versement d'un capital aux bénéficiaires, une rente d'éducation ou une rente de conjoint.

Pour le versement en capital, l'assureur règle le montant du capital garanti aux bénéficiaires lorsque le décès de l'assuré est constaté. Si le contrat est une couverture de prêt, le capital est versé à l'organisme prêteur.

La rente d'éducation ou rente des enfants est conçue pour permettre le financement des études des enfants. En cas de décès de l'assuré, elle prévoit le versement d'une rente aux bénéficiaires jusqu'à une date donnée, par exemple le 31 décembre suivant le 25^{ème} anniversaire du bénéficiaire. Il s'agit toujours d'une rente temporaire.

La rente de conjoint est versée au bénéficiaire à partir du décès de l'assuré. Son versement peut être prévu pour une période donnée (par exemple jusqu'au 31 décembre suivant le 60^{ème} anniversaire du conjoint) ou pour toute la vie du bénéficiaire.

- Incapacité

L'assuré qui se trouve dans l'incapacité physique constatée par le médecin traitant de continuer ou de reprendre le travail est déclaré comme étant en incapacité temporaire de travail, usuellement désignée par "incapacité". Suivant le contrat, la garantie incapacité peut porter sur l'incapacité au métier de l'assuré, ou à un métier quelconque. La Sécurité Sociale prévoit le versement d'indemnités journalières (IJ) pour un assuré en incapacité.

Un assuré qui se trouve, par la suite de maladie ou d'accident, dans l'impossibilité totale et médicalement justifiée d'exercer sa profession pendant une durée ininterrompue de plus de 365 jours, est dit comme étant en longue maladie. Cet état peut durer jusqu'au 1 095^{ème} jour d'arrêt de travail, si l'invalidité n'a pas été reconnue avant cette date.

- Invalidité

L'invalidité concerne les salariés qui du fait de leur état de santé présentent une capacité de gain ou de travail réduite des 2/3. L'article L341-4 du code de la Sécurité Sociale distingue trois catégories d'invalidité :

- Catégorie 1 : invalides capables d'exercer une activité rémunérée,
- Catégorie 2 : invalides absolument incapables d'exercer une profession quelconque,

- Catégorie 3 : invalides qui, étant absolument incapables d'exercer une profession, sont, en outre, dans l'obligation d'avoir recours à l'assistance d'une tierce personne pour effectuer les actes ordinaires de la vie.

2.1.3 Les produits du portefeuille

- Vie Entière (VE)

C'est une assurance qui garantit le versement de capital fixe lors du décès de l'assuré quelle que soit la date de survenance du décès.

- Capitaux différés (CD)

L'assureur s'engage, moyennant le versement d'une prime (unique ou périodique) par l'assuré, à verser un capital sous condition que l'assuré soit en vie au-delà d'une échéance fixée.

- Mixtes différées (MX)

C'est l'adjonction d'une assurance temporaire à une assurance de capital différé. L'assurance temporaire garantit le paiement d'un capital au bénéficiaire désigné si le décès de l'assuré survient avant une date déterminée. Si à l'échéance l'assuré est toujours vivant, alors l'assureur n'a plus aucune obligation.

La mixte différée porte ainsi sur deux risques contradictoires : le décès d'un côté et la survie d'un autre. Cette assurance permet donc à l'assuré de contracter dans une même police une opération de prévoyance et la constitution d'une épargne.

2.2 Présentation des bases de données

Nous utilisons dans le cadre de cette étude deux bases de données. La première est la base de données des *model points* et la seconde, construite à partir de la première est une base de données des cash flows déterministes.

2.2.1 Les *model points*.

Notre base de données initiale comporte 18 525 *model points*, soit 15% des *model points* du périmètre Epargne retraite prévoyance. Elle ne comporte pas d'affaires nouvelles. Le tableau suivant décrit le portefeuille étudié :

Nombre de contrats	Prime en €	PM en €
236 814	91 028 843	745 014 962

Tableau 1 : Description quantitative du portefeuille

2.2.1.1 Analyse univariée

Notre base de données comporte 47 variables dont 12 qualitatives et 35 variables quantitatives. Nous supprimons les variables à faible variabilité, celles qui ont une seule modalité (par exemple, Fund, Product, année d'évaluation, mois d'évaluation), celles qui sont utilisées comme identifiant ou clé (par exemple : record_no, progressive, rescale). En outre, nous supprimons les variables qui sont corrélées entre elles (par exemple les variables cnt_durée / durée_cnt ; pr_duree/duree_pp). Enfin nous supprimons les variables qui ont une faible variabilité et qui ne sont pas pertinentes du point de vue business (coefficient de validation, etc).

Il nous reste donc les 15 variables suivantes :

- Pour les variables quantitatives

Variables	Descriptions
age_sous	Age à la souscription
anc_cnt	Ancienneté du contrat
coefcup_sa	Pourcentage du capital dans la somme assurée de type prime unique
coefcus_sa	Pourcentage de la contre assurance dans la somme assurée de type prime unique
duree_cnt	Durée du contrat
duree_pp	Durée de versement des primes périodiques
fam_tit	Famille des taux d'intérêt technique
freq_pp	Fréquence de paiement des primes
nb_cnt_ini	Nombre de contrats provenant de la base de données
pm_init	PM initiale provenant des bases de données
prime_pp	Montant de prime périodique provenant de la base
sa_pp_init	Somme assurée de type prime périodique
sa_pu_init	Somme assurée de type prime unique
sous_annee	Année civile de souscription
tx_chg_acq	Taux de chargement d'acquisition issu de la base de données

Tableau 2 : Description des variables quantitatives

Nous remarquons que ces variables peuvent être regroupées en trois catégories :

- les variables liées au **temps** : âge de souscription, ancienneté du contrat, durée du contrat, durée de versement des primes périodiques
- les variables liées aux **capitaux investis** : somme assurée, provision mathématique initiale
- les variables liées aux **taux** : taux d'intérêt technique, taux de chargement d'acquisition

Ce regroupement de variables nous sera utile pour identifier les pistes de calibrage de notre algorithme d'agrégation.

- Pour les variables qualitatives

Variables	Descriptions
code_group	Code Groupe
code_prod	Code produit
code_res	Code réseau
fam_prod	Famille de produit
fam_tbl_mo	Famille de table de mortalité
fam_tit_us	Famille de taux d'intérêt
Group	Variable utilisée pour découper la base par groupe de contrats
product_it	Code produit intérêt technique
produit	Code produit

Figure 9 : Description des variables qualitatives

La variable qualitative la plus pertinente est la variable « group ». Elle est l'une des plus dispersées, possède peu de valeurs extrêmes et a une pertinence business (combinaison du code produit, du taux d'intérêt technique et du code réseau) (voir annexe pour plus de détails). C'est d'ailleurs cette variable qui est utilisée pour le découpage actuel du portefeuille.

Avec ces variables nous observons les points suivants :

- le portefeuille est constitué principalement des vies entières (73%). Les autres produits se répartissent entre 10% pour les mixtes différés et 17 % pour les capitaux différés.
- on y retrouve 39 produits souscrits entre 1925 et 2019. La figure ci-après montre la distribution du portefeuille par produit en fonction du nombre de MP.

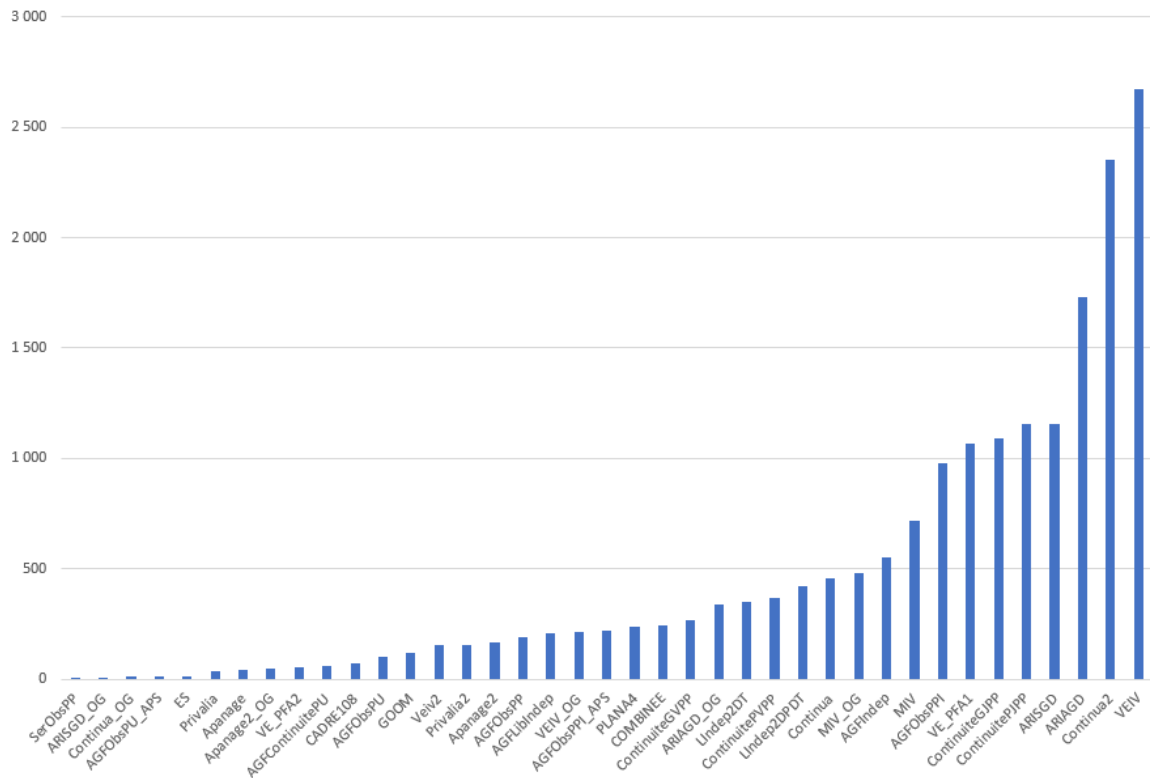


Figure 10 : Distribution du portefeuille par produit en fonction du nombre de MP

Cette distribution montre qu'il y a :

- 7 produits qui se retrouvent dans plus de 1 000 *model points*.
- 32 produits qui se retrouvent dans moins de 1 000 *model points*.

Les capitaux assurés sont en moyenne de 47 861 € par model point. La provision mathématique initiale de notre portefeuille est de 745 014 962 €. Le diagramme suivant donne sa répartition selon les 2 groupes de produits précédemment identifiés :

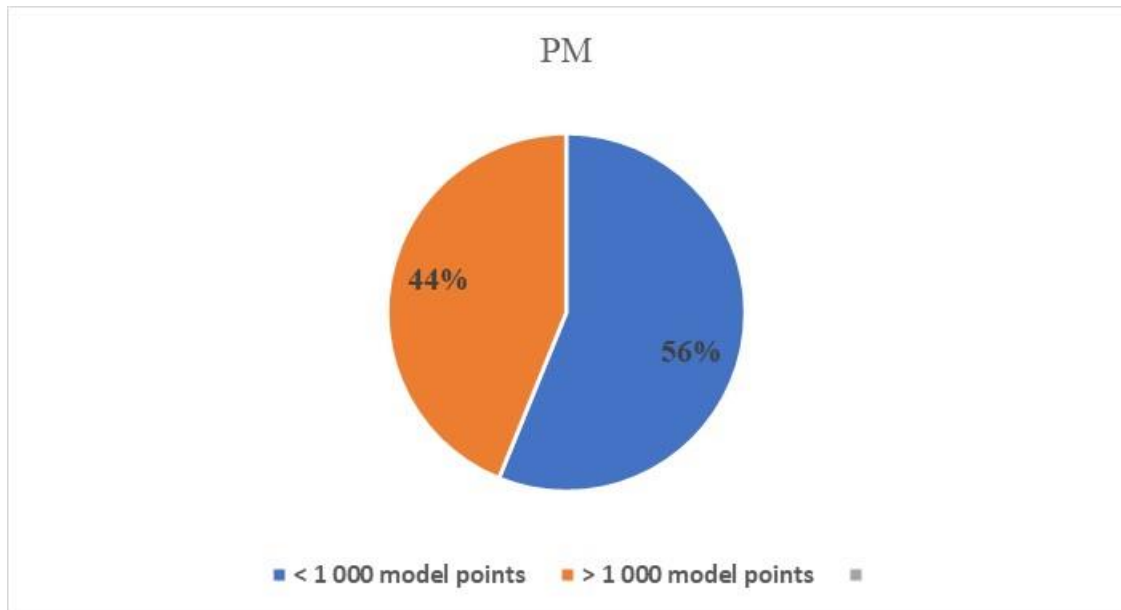


Figure 11 : Répartition des capitaux assurés en fonction du nombre de MP

On observe une répartition à peu près équitable de la PM entre le groupe de produits qu'on retrouve dans plus de 1 000 *model points* et le groupe de produits que l'on retrouve dans moins de 1 000 *model points*.

Ces produits sont distribués au travers de 4 réseaux, le principal étant le réseau AF (57%).

Les graphiques suivants donnent les informations sur l'âge de souscription.

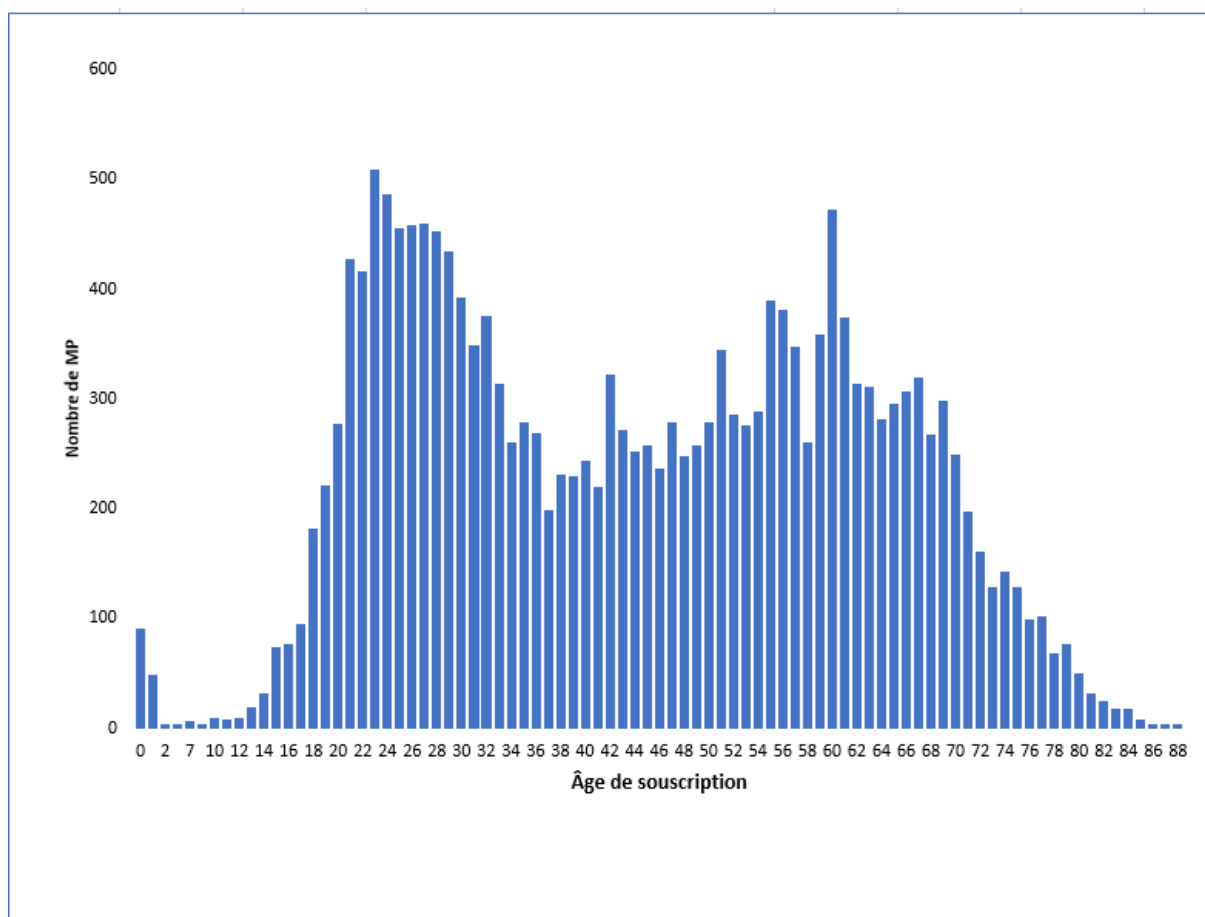


Figure 12 : Répartition des MP en fonction de l'âge de souscription des assurés.

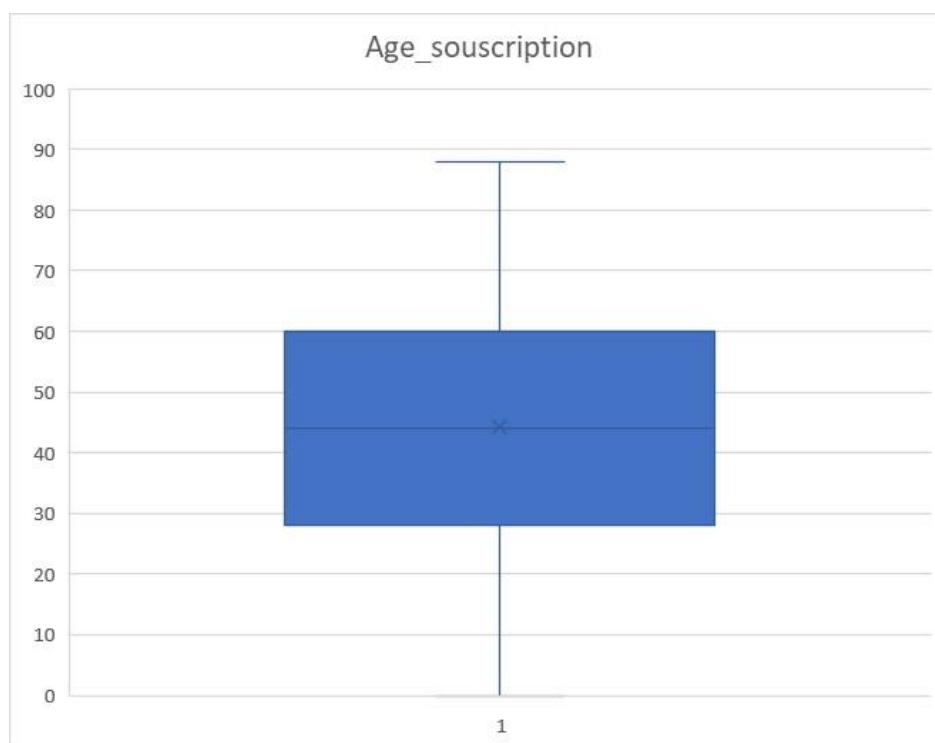


Figure 13 : Diagramme en moustache de l'âge de souscription

On remarque que l'âge moyen est de 44 ans. Cette variable est très dispersée (écart type de 17,7), avec une valeur minimale de 0 et une valeur maximale de 88.

Les taux techniques varient entre 0% et 4,50% et près de la moitié du nombre des MP (47 %) ont un taux technique de 3,50%.

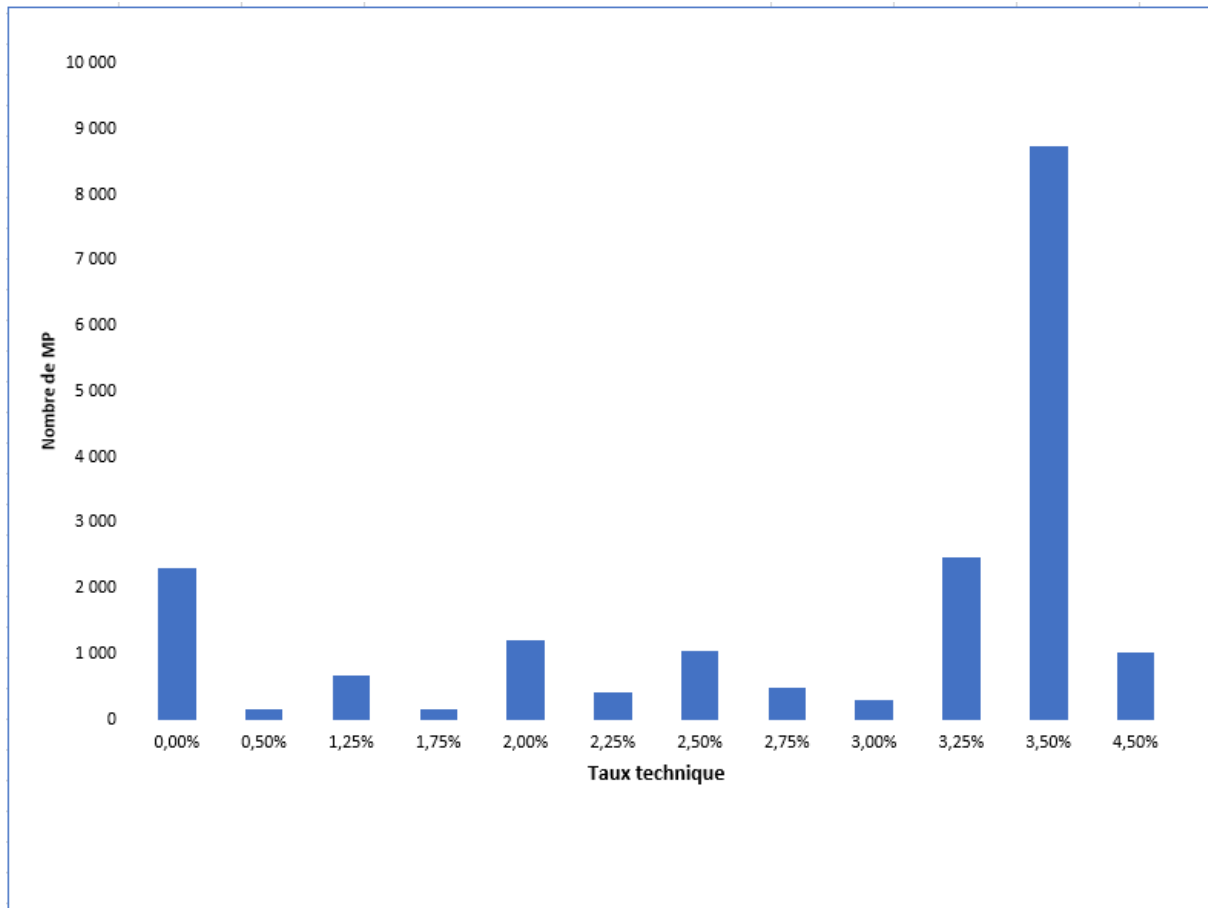


Figure 14 : Distribution des MP en fonction des taux techniques

Le complément de l'analyse descriptive se trouve en annexe.

2.2.1.2 Analyse bivariée

Dans la suite de l'étude nous allons nous intéresser uniquement aux variables quantitatives. Pour cette raison, l'analyse bivariée se fera uniquement sur ces variables. La matrice des corrélations nous permet d'identifier les liens entre les variables quantitatives.

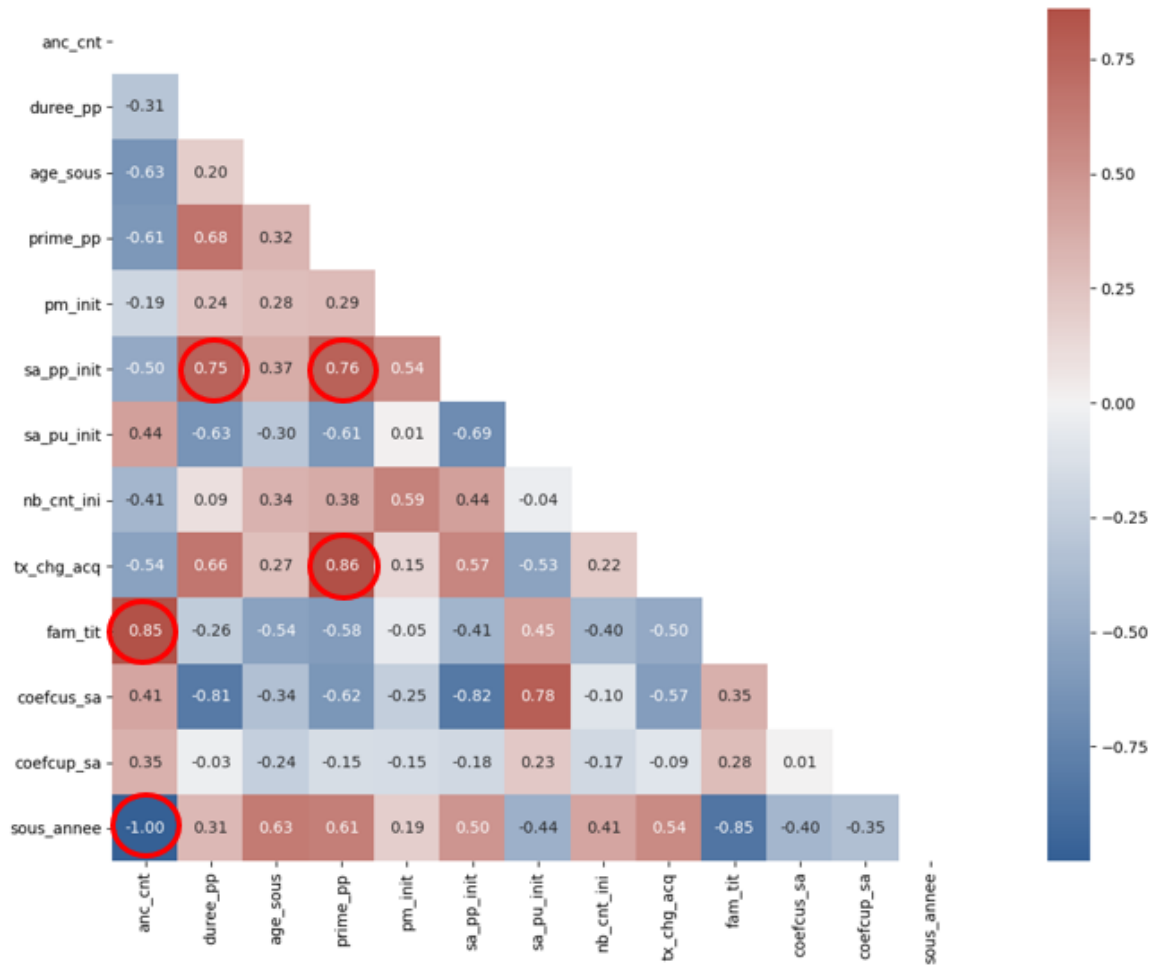


Figure 15 : Matrice de corrélation des variables quantitatives

On peut remarquer que :

- l'ancienneté et l'année de souscription sont parfaitement corrélées,
- le taux de chargement d'acquisition et les primes périodiques sont fortement corrélés (0.86). Il en est de même :
 - o de l'ancienneté et du taux d'intérêt technique (0.85),
 - o du montant des capitaux investis en prime périodique et le montant de la prime périodique (0.76),
 - o et du montant des capitaux investis en prime périodique et la durée de versement des primes périodiques (0.75).

2.2.2 Base de données des cash flows déterministes

A partir de la première base de données, nous avons construit une seconde base de données en suivant les étapes suivantes :

- Nous avons supprimé le regroupement qualitatif actuel des MP (combinaison du code produit, du taux d'intérêt technique et du code réseau). Chacune des 18 525 lignes de la base précédente constitue ainsi un groupe de MP et les calculs se feront désormais ligne par ligne.
- Avec 18 525 lignes les temps de calcul seront longs. Nous avons donc découpé notre base de données en trois sous bases selon le type de produit de notre portefeuille (vie entière, capitaux différés, mixtes différés).
- Nous avons ensuite déterminé les cash flows déterministes sur chacun de ces périmètres. Les résultats ont été regroupés dans une seconde base de données.
- Cette seconde base comporte 305 variables calculées à chaque pas de temps (0 à 60). Après un nettoyage de cette base de données et une procédure de sélection des variables (voir Annexe 1 : Statistiques descriptives - variables quantitatives - sélection des variables), nous avons supprimé 86 variables pour obtenir une base avec des variables pertinentes pour notre étude.

Le schéma ci-après présente ces différentes étapes.

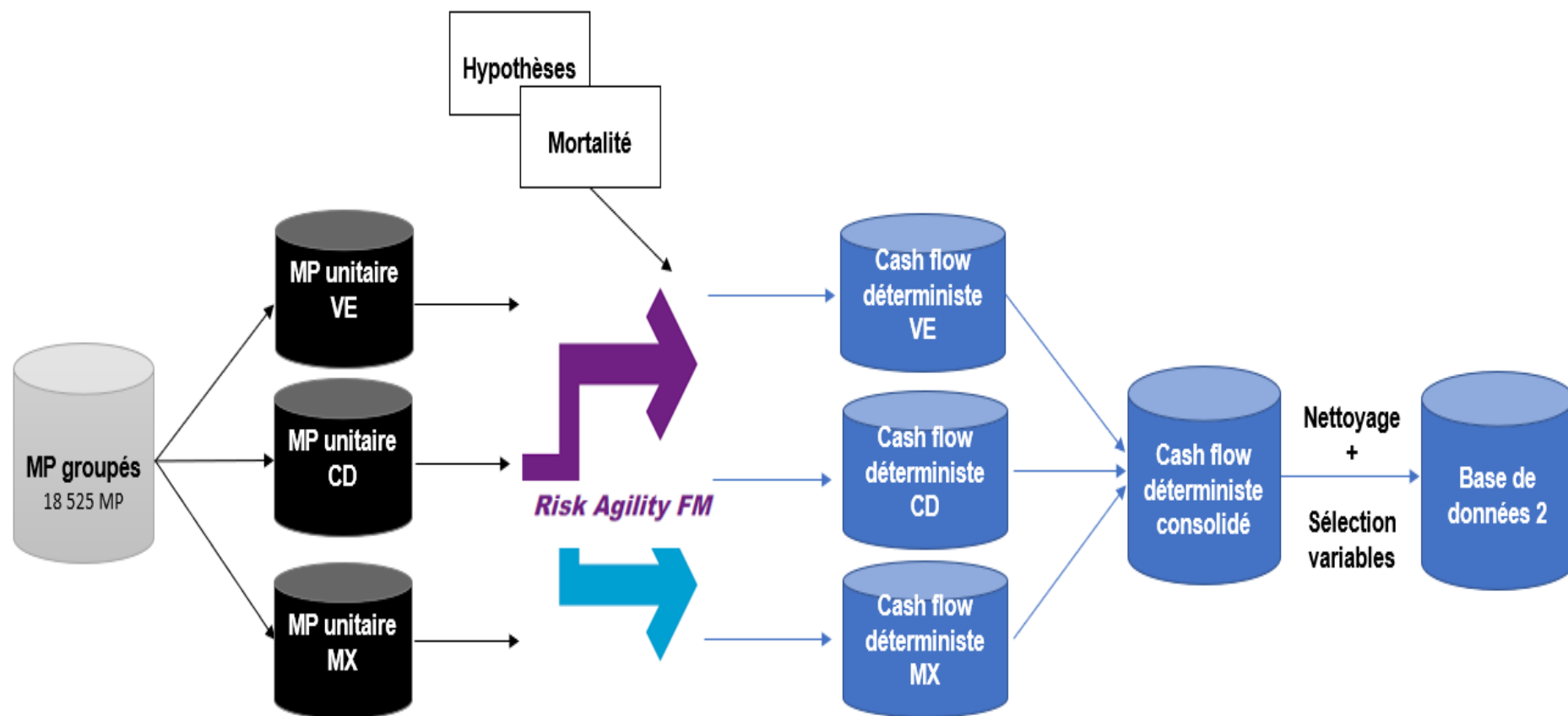


Figure 16 : Construction de la seconde base de données

Avec cette nouvelle base de données, nous pouvons obtenir d'autres statistiques descriptives.

2.2.2.1 Contribution des garanties au BEL

Le tableau ci-après donne la contribution de chacune des garanties de notre portefeuille, capitaux décès (CD), mixtes différés et vie entière (VE) au BEL total.

Garanties	Contribution au BEL (%)
CD	9,06
MX	8,98
VE	81,95
Total général	100,00%

Tableau 3: Contribution des différentes garanties au BEL

La vie entière contribue majoritaire au BEL de notre portefeuille.

2.2.2.2 Contribution des MP au BEL

La figure suivante donne le BEL par MP en euros.

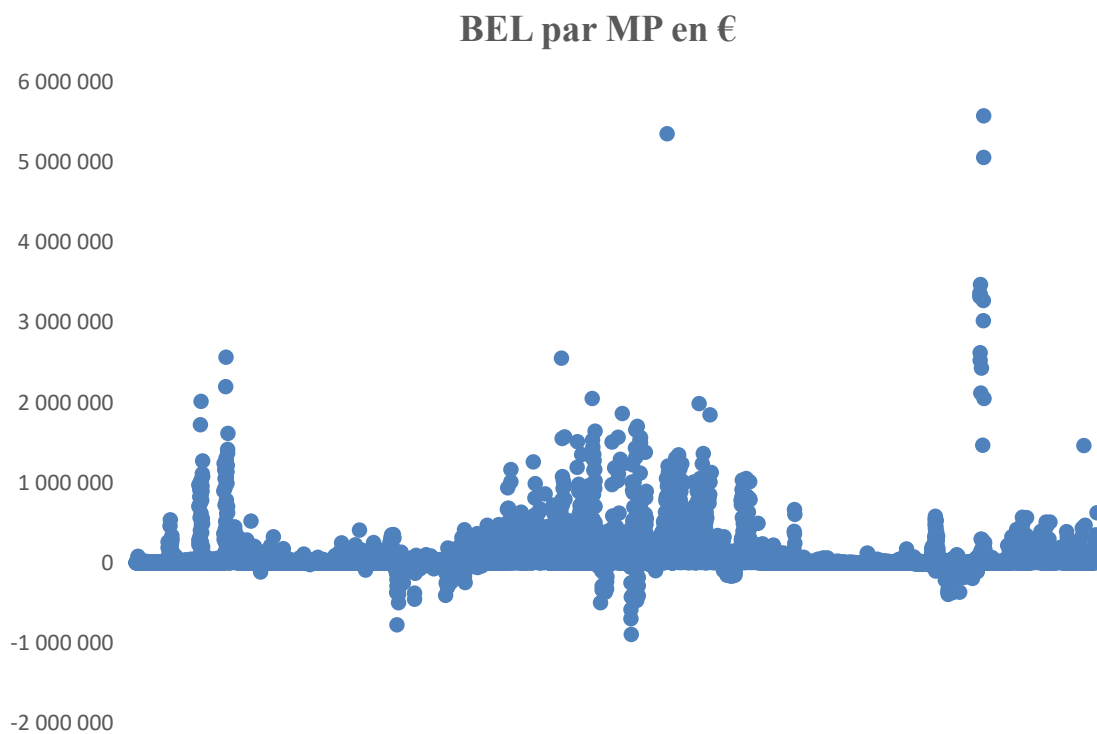


Figure 17: Contribution des MP au BEL initial

Nous pouvons faire les observations suivantes :

- La moyenne de BEL par MP est de 44 K€ avec un écart type de 168 K€.
- Environ 1500 MP sur les 18 500 MP, soit moins de 10%, ont un BEL négatif. Ces MP représentent un BEL de -37 M€ sur les 818 M€ du portefeuille, soit moins de 5%. Effectuer le calibrage en utilisant la valeur absolue ne changera pas significativement les résultats.

Ce chapitre nous a permis de comprendre les produits de notre portefeuille et de les décrire. Ce qui nous permettra de mieux les agréger pour calibrer notre algorithme. Avant cela, il est important de bien comprendre le modèle d'agrégation des données.

3. MODELE D'AGREGATION DES DONNEES

L'agrégation des *model points* se fait actuellement via un algorithme appelé le LOP. Cet algorithme repose sur les moindres carrés non négatifs et est déroulé suivant une procédure spécifique dans le modèle *Full ALM*. Dans ce chapitre, nous présenterons dans une première partie les aspects théoriques des moindres carrés non négatifs, *Non-Negative Least Squares* (NNLS), puis dans une seconde partie son implémentation chez Allianz. Nous terminerons en présentant les résultats obtenus actuellement sur le périmètre de la prévoyance individuelle.

3.1 *Non-Negative Least Squares* (NNLS)

3.1.1 Moindres carrés

Les moindres carrés sont une méthode d'approximation utilisée pour estimer les solutions d'un système d'équations. Le système peut contenir plus d'équations que de variables (sur déterminé) ou plus de variables que d'équations (sous déterminé). Les moindres carrés sont classés en deux groupes, les moindres carrés linéaires et les moindres carrés non linéaires. Les moindres carrés linéaires sont souvent de formes fermées, c'est à dire qu'ils peuvent être résolus par un nombre fini d'opérations standards. Voici un exemple d'ajustement utilisant les moindres carrés.

Soit t une variable indépendante et soit $b(t)$ une fonction inconnue de t qu'on souhaite estimer. Supposons que nous avons m observations.

$$b_i = b(t_i), \quad i = 1, \dots, m.$$

L'objectif est de construire $b(t)$ comme combinaison linéaire de fonctions de base, $\phi_j(t)$, qui ne sont pas des fonctions linéaires de t :

$$b(t) \approx A_1\phi_1(t) + \dots + A_n\phi_n(t) \quad (3.1)$$

La matrice obtenue, que nous nommons x , est une matrice rectangulaire de dimension $m \times n$. Elle est définie comme suit :

$$x_{i,j} = \phi_j(t_i) \quad i = 1, \dots, m \quad \text{et } j = 1, \dots, n \quad \text{ss}$$

On peut écrire l'équation (3.1) sous forme matricielle

$$b \approx Ax. \quad (3.2)$$

Les variables inconnues A_j sont linéaires dans le modèle. Le système d'équation linéaire (3.2) est sur déterminé, c'est à dire qu'il a plus d'équations que de variables qu'on souhaite estimer. Les résidus sont la différence entre les valeurs observées et les valeurs obtenues du modèle. Mathématiquement, on peut écrire les résidus ainsi :

$$r_i = b_i - \sum_j A_j \phi_j(t) \quad i = 1, \dots, m$$

Nous cherchons A_j qui donne la plus petite somme des carrés des résidus. En d'autres termes, les moindres carrés minimisent la somme des carrés des résidus :

$$\min \|r^2\| = \min \sum_{i=1}^m r_i^2$$

Les moindres carrés non linéaires peuvent être formulés comme suit :

$$F(X) = \frac{1}{2} \sum_{i=1}^m (f_i(x))^2$$

Où f_1, \dots, f_m sont des fonctions de \mathbb{R}^n vers \mathbb{R} . Les moindres carrés non linéaires n'ont pas de solution fermée et doivent être résolus numériquement.

Pour pouvoir utiliser les moindres carrés non négatifs (NNLS), nous avons besoin de certaines hypothèses sur les matrices de notre base de données. Les matrices doivent être **positives**, **irréductibles** et doivent satisfaire aux conditions du **théorème de Perron-Frobenius**. Toutes les matrices utilisées ont des nombres réels.

3.1.1.1 Théorème de Perron-Frobenius

Définissons les concepts qui nous aideront à formuler ce théorème et à le démontrer.

Définition 1

Soit A une matrice. A est positive si toutes ses composantes $a_{i,j} \geq 0$ pour chaque i et j . On écrit $A \geq 0$. Il s'ensuit que sa transposée, A^T , est aussi une matrice positive et le système d'équations linéaire $Ax = b$ a des solutions positives si $b \geq 0$.

Définition 2

Soit A une matrice carrée $n \times n$. Un scalaire λ dans \mathbb{R} est appelé valeur propre de A , si l'une des conditions équivalentes suivantes est vérifiée :

- $\text{Ker}(A - \lambda I_n) \neq \{0\}$, ou I est la matrice unitaire.
- $\text{Det}(A - \lambda I_n) = 0$,
- Il existe un vecteur non nul x de \mathbb{R}^n , solution de l'équation $Ax = \lambda x$. Le vecteur x est dit vecteur propre de A associé à la valeur propre λ .

Définition 3

L'ensemble des valeurs propres de A dans \mathbb{R} est appelé spectre sur \mathbb{R} de la matrice A et sera noté dans la suite $S_p(A)$.

Définition 4

Soit A une matrice carrée $n \times n$. Une valeur propre λ de A est dite dominante si, pour tout $\lambda' \in S_p(A) - \{\lambda\}$, on a $|\lambda| > |\lambda'|$. Le rayon spectral de A , noté $\rho(A)$ est le plus grand des modules des valeurs propres de A :

$$\rho(A) = \max \{|\lambda| \mid \lambda \in S_p(A)\}$$

Définition 5

Une matrice carrée positive T est dite primitive s'il existe $k > 0$ tel que $T^k > 0$, c'est-à-dire que $T_{i,j}^{(k)} > 0$ pour tout i, j .

Définition 6

Une matrice A $n \times n$ est dite réductible si pour $n \geq 2$ et si et seulement si, il existe une matrice de permutation P telle que

$$PAP^T = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$$

Avec B et D des matrices carrées et 0 , la matrice 0 . La matrice A est irréductible si elle n'est pas réductible.

Remarque 1

Une matrice A est réductible si, et seulement si, sa transposé A^T est réductible.

Remarque 2

Si T est irréductible alors $I + T$ est une primitive, où I est la matrice unitaire. En effet le développement binomial montre que plus k est grand, plus on aura des coefficients positifs et donc $(I + T)^k > 0$

$$(I + T)^k = I + kT + \frac{k(k-1)}{2}T^2 + \dots$$

Ces définitions posées, nous pouvons maintenant énoncer le théorème de Perron-Frobenius.

Théorème de Perron-Frobenius

Soit A une matrice $n \times n$ positive irréductible

1. $\rho(A)$ est une valeur propre de A ,
2. $\rho(A) > 0$,
3. $\rho(A)$ est simple,
4. Il existe un vecteur propre $x > 0$, tel que $Ax = \rho(A) x$
5. Les seuls vecteurs propres positifs de A sont les multiples positifs du vecteur de Perron P , où P est l'unique vecteur propre strictement positif P associé à la valeur propre $\rho(A)$, tel que $\|P\| = 1$.

En plus du théorème de Perron-Frobenius, nous avons besoin des conditions de Karush-Kuhn-Tucker (KKT), qui sont une généralisation du multiplicateur de Lagrange.

3.1.1.2 Conditions de Karush-Kuhn-Tucker

Les conditions de KKT doivent être satisfaites pour obtenir une solution optimale. Dans notre cas, il s'agit de la solution ayant la plus petite somme du carré des résidus. Cette condition est nécessaire pour que la solution soit optimale, mais pas suffisante pour montrer que la solution est optimale.

Soit f une fonction objective, h et g les fonctions restrictives (contraintes). Soit x_{min} le minimum local de f que nous recherchons.

$$\min_x f(x) \text{ sous contrainte de } \begin{cases} h(x) = 0 \\ g(x) \leq 0 \end{cases}$$

Il existe λ et μ tel que

$$\nabla f(x) + \sum_{i=1}^m \mu_i \nabla g_i(x) + \sum_{j=1}^l \lambda_j \nabla h_j(x) = 0 \quad (3.3)$$

$$g_i(x) \leq 0 \quad \forall i \in \{1 \dots m\} \quad (3.4)$$

$$h_j(x) = 0 \quad \forall j \in \{1 \dots l\}$$

$$\mu_i \geq 0 \quad \forall i \in \{1 \dots m\} \quad (3.5)$$

$$\mu_i g_i(x) = 0 \quad \forall i \in \{1 \dots m\}$$

3.1.2 La méthode des moindres carrés non négatifs

Un problème fondamental en modélisation est l'estimation des paramètres d'un modèle pour décrire des données. Nous supposons que nous avons plusieurs observations qui sont des fonctions linéaires des paramètres. Ayant un grand nombre de telles observations, nous pouvons estimer les paramètres du modèle. Soit $x = (x_1, \dots, x_n)^T$ les paramètres inconnus du modèle. Les différentes expériences relatives à x sont représentées par $A \in \mathbb{R}^{m \times n}$, et les valeurs observées par $b \in \mathbb{R}^m$. L'objectif est de construire un vecteur $x \in \mathbb{R}^n$ qui explique les valeurs observées le mieux possible. Cela peut s'écrire ainsi

$$Ax = b$$

Ce système peut être sous déterminé ($m < n$) ou sur déterminé ($m \geq n$). Dans le dernier cas, la méthode des moindres carrés permet de déterminer x tel que la somme des résidus des moindres carrés est minimisée :

$$f(x) = \frac{1}{2} \|Ax - b\|^2 \quad (3.6)$$

Cependant, dans de nombreux problèmes, les paramètres représentent des quantités qui ne peuvent prendre que des valeurs positives. Dans ce cas, le problème (3.6) doit être modifié pour inclure des contraintes de non-négativité sur les paramètres du modèle x . Le problème qui en résulte est appelé les moindres carrés non négatifs (NNLS) et est formulé comme suit :

Étant donnée une matrice $A \in \mathbb{R}^{m \times n}$, et l'ensemble des valeurs observées donné par $b \in \mathbb{R}^m$, trouver un vecteur positif $x \in \mathbb{R}^n$ qui minimise la fonction $f(x) = \frac{1}{2} \|Ax - b\|^2$. C'est-à-dire

$$\min_x f(x) = \frac{1}{2} \|Ax - b\|^2, \quad x \geq 0 \quad (3.7)$$

Le gradient de $f(x)$ est

$$\nabla f(x) = A^T(Ax - b)$$

et les conditions d'optimalité de KKT pour le problème de NNLS sont :

$$\begin{aligned} x &\geq 0 \\ \nabla f(x) &\geq 0 \\ \nabla f(x)^T x &= 0 \end{aligned}$$

Certaines des méthodes itératives de résolution de ce problème sont basées sur la complémentarité linéaire (LCP). Vérifions que les contraintes ci-dessus remplissent les conditions de KKT.

Des conditions de KKT on peut écrire que :

$$\begin{aligned} f(x) &= \frac{1}{2} \|Ax - b\|^2 \\ g(x) &= -x \\ h(x) &= 0 \end{aligned}$$

D'où

$$\begin{aligned} \nabla f(x) &= A^T(Ax - b) \\ \nabla g(x) &= -1 \end{aligned}$$

La première contrainte ressort clairement de la définition du problème (3.7). La deuxième et troisième contraintes sont équivalentes si on les substitue en (3.3), (3.4) et (3.5) et on obtient :

$$\begin{aligned} \nabla f(x) - \mu^T &= 0 \\ \mu^T x &= 0 \\ \mu &\geq 0 \end{aligned}$$

On peut ainsi remarquer que :

$$\nabla f(x) = \mu^T \geq 0 \quad \Rightarrow \quad \nabla f(x) \geq 0$$

$$-\nabla f(x)^T x = 0 \quad \Rightarrow \quad \nabla f(x)^T x = 0$$

Problème de complémentarité linéaire :

Étant donnée une matrice $A \in \mathbb{R}^{m \times n}$, et l'ensemble des valeurs observées donné par $b \in \mathbb{R}^m$, trouver un vecteur positif $x \in \mathbb{R}^n$ qui minimise la fonction

$$\begin{aligned} \lambda = \nabla f(x) = A^T(Ax - b) &\geq 0 \\ x &\geq 0 \\ \lambda^T x &= 0 \end{aligned}$$

Ce problème est essentiellement l'ensemble des conditions d'optimalité KKT pour la programmation quadratique. Il se réduit à trouver un x positif qui vérifie

$$(Ax - b)^T Ax = 0$$

Il est difficile de gérer les contraintes non négatives avec plusieurs équations non linéaires. Une formulation différente de NNLS est parfois plus gérable en utilisant la variable vectorielle résiduelle :

$$\begin{aligned} P &= b - Ax \\ \text{avec } Ax + P &= b \\ \text{et } x &\geq 0 \end{aligned}$$

L'avantage de cette formulation est que nous avons une fonction objective simple et séparable avec des contraintes linéaires et non négatives.

3.1.3 L'algorithme NNLS

L'algorithme NNLS détermine la solution sous la forme d'un vecteur x . Ce vecteur résout le problème des moindres carrés, où $\|Ax - b\|$ est à minimiser avec la contrainte $Gx \geq h$. A est une matrice $m \times n$ et b est un vecteur avec m éléments.

Ce qui distingue le NNLS des autres optimisations des moindres carrés, c'est que A dans d'autres optimisations des moindres carrés a tendance à être une matrice arbitraire $m \times n$, mais

dans NNLS, G est la matrice d'identité et h est le vecteur nul. Ce problème a toujours au moins une solution, et cette solution contiendra généralement au moins $n - m$ éléments égaux à zéro.

3.2 L'algorithme NNLS chez Allianz

3.2.1 Approche méthodologique

L'algorithme NNLS, appelé LOP (Linear OPTimization) par la suite, peut être appliqué afin d'agréger un grand volume de données, qui sont dans notre cas les *model points*. Le but est de créer une table de *model points* (MP) significativement plus petite que la table de MP initiale, sans que cela n'altère les résultats de la projection (en particulier sur les variables clefs : BEL, PVFP). L'algorithme LOP n'agrège pas les *model points*, mais les remplace par des *model points* représentatifs, déjà présents dans le portefeuille, pondérés, pour minimiser la différence entre des valeurs cibles calculées *model point* par *model point* et ces mêmes valeurs cibles recalculées après pondération.

Concrètement, le LOP consiste à trouver les poids x_i positifs solutions de l'équation

$$Ax = b$$

où :

- A est une matrice $n \times m$, n est le nombre de paramètres dans la fonction cible et m est le nombre de *Model Points* (MP).
- b est un vecteur de n éléments, représentant l'agrégation des paramètres cible pour l'ensemble des MP.
- x est le vecteur poids résultant.

L'équation précédente peut aussi s'écrire de façon détaillée comme suit :

$$x_1 \times \begin{pmatrix} Res_1(1) \\ Res_1(2) \\ \dots \\ Res_1(n) \\ \dots \\ CF_1(1) \\ CF_1(2) \\ \dots \\ CF_1(n) \end{pmatrix} + x_2 \times \begin{pmatrix} Res_2(1) \\ Res_2(2) \\ \dots \\ Res_2(n) \\ \dots \\ CF_2(1) \\ CF_2(2) \\ \dots \\ CF_2(n) \end{pmatrix} + \dots + x_m \times \begin{pmatrix} Res_m(1) \\ Res_m(2) \\ \dots \\ Res_m(n) \\ \dots \\ CF_m(1) \\ CF_m(2) \\ \dots \\ CF_m(n) \end{pmatrix} = \begin{pmatrix} Res_G(1) \\ Res_G(2) \\ \dots \\ Res_G(n) \\ \dots \\ CF_G(1) \\ CF_G(2) \\ \dots \\ CF_G(n) \end{pmatrix}$$

avec

- m le nombre de lignes de la matrice des *model points* avant LOP
- n la maturité de projection des *model points*
- $\frac{Res_i(k)}{CF_i(k)}$: Grandeur du model point i , réserves (*Res*) ou cash-flows (*CF*), projetée à la date k (constituant la matrice A)
- $\frac{Res_G(k)}{CF_G(k)}$: Grandeur totale du portefeuille de base par année à la date de projection k (constituant le vecteur b)

Le LOP s'applique après une première segmentation du portefeuille en sous-groupes homogènes. L'algorithme s'applique par sous-groupe. En effet, pour une exécution rapide de l'algorithme, les calculs sont effectués non pas sur toute la base de données des MP (18 525 lignes), mais plutôt sur la base de sous-groupes (185 sur le portefeuille étudié). Actuellement, les sous-groupes sont créés sur la base de la variable « group » des MP (concaténation du code produit, du taux d'intérêt technique et du réseau de distribution). Cette segmentation crée obligatoirement un biais dans l'agrégation, puisque les données appartenant à 2 sous-groupes différents, ne peuvent plus être regroupées.

3.2.2 Processus LOP dans le modèle LSIM

Le processus d'agrégation des MP avec l'approche LOP dans le modèle LSIM se fait en quatre étapes. Le schéma suivant présente ces différentes étapes.

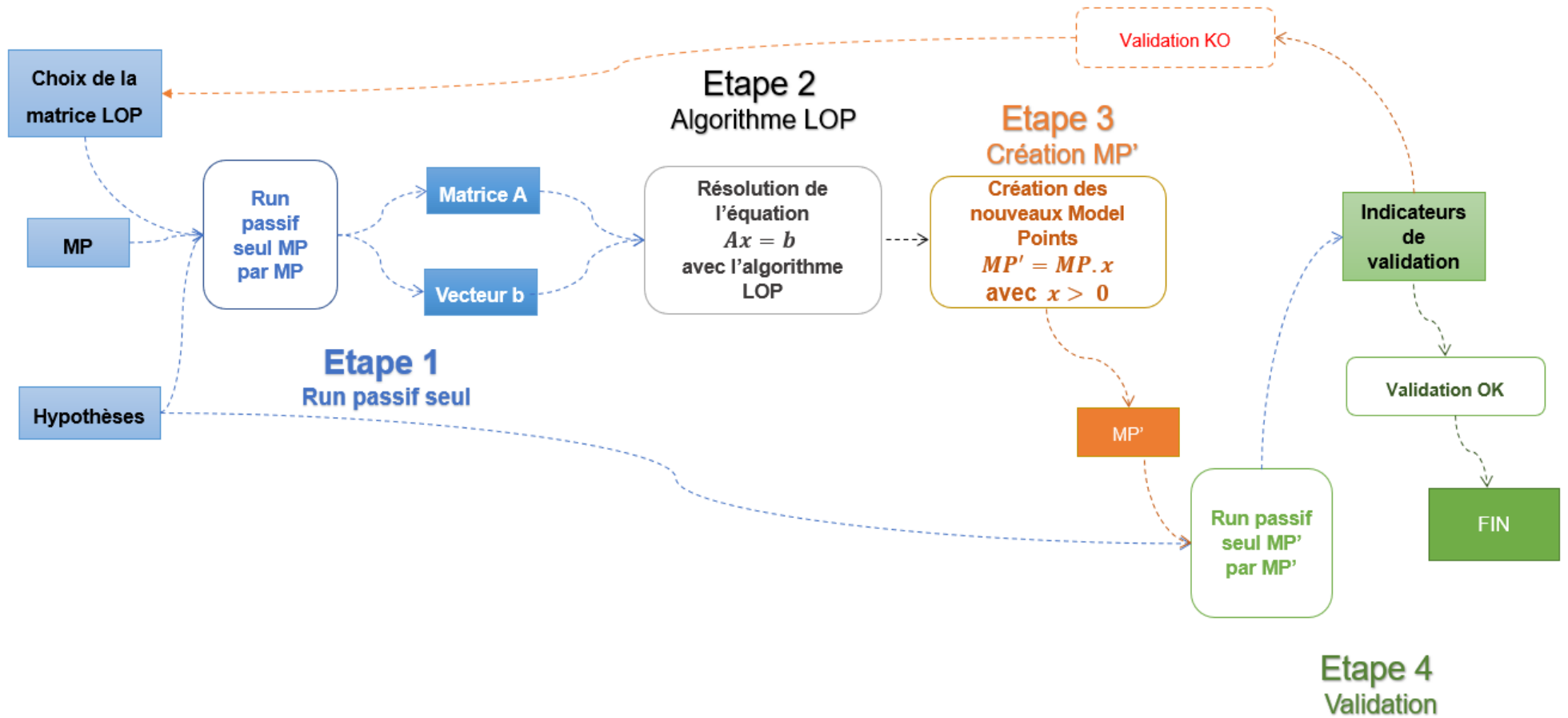


Figure 18 : Processus LOP

Etape 1 : lancement du run passif seul

Il s'agit d'un scenario best estimate MP par MP. Le lancement de ce run nécessite les inputs suivants

- Les hypothèses économiques, financières et techniques.
- Les *model points*.
- La matrice utilisée dans le cadre du LOP

Les outputs de cette étape sont :

- Le vecteur b par groupe de contrats
- Les sous matrices A par groupe de contrats

Etape 2 : Algorithme LOP

L'algorithme NNLS est déclenché à cette étape. Pour chaque groupe de MP, la sous matrice A et le vecteur b sont lus. Ensuite, l'équation $Ax = b$ est résolue, x correspond au nombre de MP pondérés.

Etape 3 : Création des nouveaux MP

Les MP sont pondérés avec les vecteurs x pour créer les nouveaux MP. Seuls les éléments avec un $x > 0$ sont inclus dans les *model points* finaux. Les poids nuls ne sont pas considérés. Les MP avec un coefficient nul sont supprimés.

Etape 4 : Validation du modèle

Le modèle en passif seul est relancé avec les nouveaux MP (MP agrégés). Les indicateurs de validation sont calculés (BEL, PVFP, etc). L'erreur d'estimation est calculée comme suit :

$$\varepsilon = \frac{\text{Indicateur}_{MP \text{ Agrégés}} - \text{Indicateur}_{MP/MP}}{\text{Indicateur}_{MP/MP}}$$

Si l'erreur est acceptable, le processus est terminé. Sinon, un nouveau run peut être lancé avec des nouvelles variables LOP (Etape 1). La qualité des résultats ne dépend pas uniquement de cette erreur. Il faut aussi apprécier la qualité de la compression des *model points*, c'est à dire le

nombre de MP après compression. La taille des MP après compression déterminera la durée du run stochastique.

3.3 Applications sur le périmètre prévoyance individuelle

Sur le périmètre prévoyance individuelle, dans la version actuelle de l'algorithme, la variable (matrice) de paramétrage du LOP est la Provision Mathématique (PM). L'indicateur d'appréciation de la qualité d'agrégation choisi est le Best Estimate Liability (BEL).

Avec une base initiale avant agrégation de 18 525 MP, regroupée en 185 groupes selon la variable « group », l'algorithme a permis d'obtenir à $t = 0$, 4 056 MP, soit une réduction de 78%. En outre, le BEL initial est de 818 Millions €. Le BEL après agrégation est de 787 Millions €, soit une erreur d'estimation de 3,78%.

Ce chapitre nous a permis de comprendre le fonctionnement de l'algorithme LOP. C'est un algorithme qui ressemble à celui des moindres carrés avec pour input les cash flows de passif calculés selon un scénario déterministe. Ces cash flows sont calculés sur la base d'une matrice, la variable LOP. Les outputs de l'algorithme sont des coefficients positifs ou nuls qui serviront de clé d'agrégation des *model points* : les coefficients nuls permettent d'annuler les lignes de *model points*, tandis que les coefficients positifs permettent le regroupement.

Le processus LOP quant à lui comporte des inputs (Etape 1), l'algorithme (Etape 2) et les outputs (Etape 3) qui sont par la suite validés (Etape 4). S'il faut améliorer les résultats sans changer d'algorithme, il faut nécessairement optimiser les inputs (Etape 1). Optimiser les inputs revient à bien choisir la matrice LOP et à bien calibrer le run passif seul MP après MP, c'est ce que nous allons faire dans le chapitre suivant.

4. CALIBRAGE DU MODELE

L'un des principaux leviers d'optimisation du modèle d'agrégation est la matrice LOP utilisée pour calculer le run déterministe, input de l'algorithme (**calibrage 1**). Il existe un autre levier qui est le critère de regroupement des *model points* de base utilisé pour calculer ce run déterministe. En effet, le run déterministe, input de l'algorithme, est calculé lui-même sur la base d'un critère d'agrégation des *model points* de base, le groupe de contrats actuellement. Dans ce chapitre, nous allons explorer ces deux pistes pour calibrer le LOP. Nous allons nous servir notamment d'une technique de classification non supervisée, le K-Means (**calibrage 2a**), et d'une technique pour réduire les dimensions d'une base de données, l'ACP (**calibrage 2b**). Le schéma suivant permet de situer les différents calibrages dans le processus LOP.

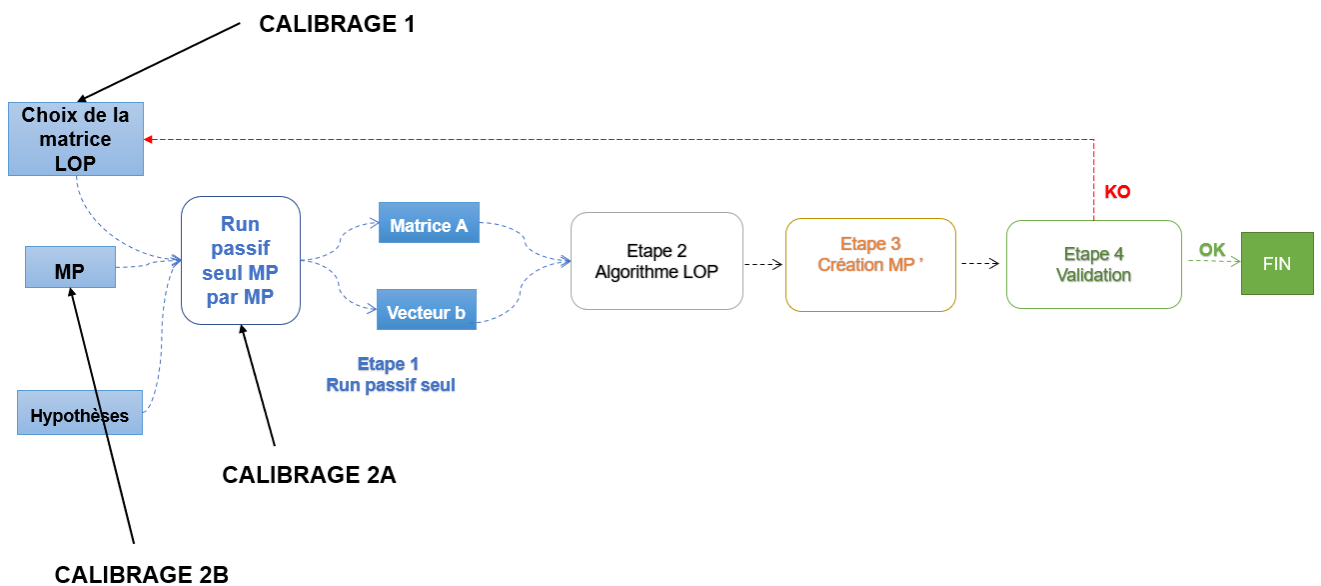


Figure 19 : Calibrage dans le processus LOP

4.1 Calibrage de la matrice LOP

Dans le processus LOP décrit au 3.2.2, le premier run est déterminé en fonction d'une matrice. Cette matrice regroupe les variables qui serviront de base de calcul de ce premier run. Elle est

de dimension 61, ce qui correspond au pas de calcul des cash flows. Actuellement, sur le périmètre prévoyance individuelle, la PM est la seule variable utilisée sur les 61 pas de temps pour cette matrice. Le choix de cette variable s'est effectué dans l'entreprise sur la base d'un jugement d'expert. Nous voulons dans cette partie proposer un calibrage plus pertinent de cette matrice.

Avec la base de données des cash flows déterministes construite au 2.2.1, nous pouvons déterminer cette matrice. Rappelons que cette base comporte 219 variables. Nous y avons ajouté la variable PVFP telle que décrite au 4.1.2.

Un programme développé sous Python nous a permis, pour chaque pas de temps de :

- calculer les corrélations entre la variable cible et les autres variables
- identifier la variable qui est la plus corrélée à la variable cible.

La variable cible est déterminée en fonction du critère d'évaluation de la qualité de l'agrégation. Nous avons retenu deux critères d'évaluation, l'erreur par rapport au BEL et l'erreur par rapport à la PVFP. Nous allons donc proposer trois matrices LOP

- la matrice LOP avec pour fonction cible le BEL,
- la matrice LOP avec pour fonction cible la PVFP,
- la matrice LOP avec pour fonction cible le BEL et la PVFP.

4.1.1 Calibrage par rapport au BEL

Le calcul des corrélations entre les variables de la base de données des cash flows déterministes et le BEL (voir Annexe 2 Choix de la matrice LOP-Corrélation au BEL pour le détail des résultats) permet d'obtenir les variables indiquées dans le tableau ci-après pour la détermination de la matrice LOP.

Pas de temps	Variable	Description
0	van_pm_clo	VAN de PM de clôture
1-7	pm_ouvbis	PM ouverture
8-11	int_ms	Intérêt sur la marge de solvabilité
12-50	ms	Marge de solvabilité
51-55	chg_ges_cu	Chargement de gestion
56-60	ms	Marge de solvabilité

Tableau 4 : Variables corrélées au BEL

4.1.2 Calibrage par rapport à la PVFP

Rappelons avant de présenter les résultats, la définition de la PVFP.

La PVFP (Présent Value of Future Profits) représente la valeur actuelle des résultats futurs d'une compagnie d'assurance.

$$PVFP = \sum_{i=1}^N \frac{R_i}{(1 + t_{xref})^i}$$

Avec

- R_i : le résultat de l'année i
- t_{xref} : le taux sans risque en année i (taux issu de la courbe swap en pratique) qui peut toutefois, sous justification, faire l'objet d'ajustement comme la prime d'illiquidité. On parlera ainsi plus de taux de référence que de taux sans risque.

Le résultat d'une compagnie d'assurance :

$$\text{Résultat}_{\text{avant impôts}} = \text{Résultat financier} + \text{Résultat technique}$$

Avec :

- Le résultat financier correspondant aux bénéfices/pertes générés par la différence entre les produits financiers réalisés par des actifs et les montants financiers crédités aux assurés (garantis et PB discrétionnaire). Ce résultat représente la marge financière que réalise l'assureur entre ses investissements et ses engagements

- Le résultat technique correspond aux bénéfices/pertes générés par la différence entre les chargements prélevés par l'assureur et ses prestations et coûts (rachats, décès, frais, commissions...). Ce résultat représente la marge technique que réalise l'assureur en raison d'une tarification et d'un provisionnement prudents.

Schématiquement le résultat de l'assureur peut se résumer à :

$$\text{Résultat}_{\text{avant impôts}} = \text{Primes} - \text{Prestations} - \text{Frais} - \text{Commissions} - \Delta PT + \text{Produits financiers}$$

Avec :

- Prestations : Rachats, décès, annuités, maturités, etc.
- ΔPT : Variation des provisions techniques (PM, PPE, RDC, PRE, ...) telle que $\Delta PT = PT_t - PT_{t-1}$
- Les produits financiers sont définis tels que :

$$\text{Produits financiers} = \text{Coupons} + \text{Dividendes} + \text{Intérêts sur la trésorerie} \pm \text{value réalisées}$$

Pour obtenir le résultat après impôts (taxes), il suffit de déduire les taxes appliquées sur les bénéfices de la compagnie :

$$\text{Résultat}_{\text{après impôts}} = \text{Résultat}_{\text{avant impôts}}(1 - t_x \text{ tax})$$

Le taux d'impôt étant identique sur toutes les années de projection, par simplification, nous utiliserons dans nos calculs le résultat avant impôts pour le calcul de la PVFP.

Le calcul des corrélations entre les variables de la base de données des cash flows déterministes et la PVFP (voir Annexe 2 Choix de la matrice LOP - Corrélation à la PVFP pour le détail des résultats) permet d'obtenir les variables indiquées dans le tableau ci-après pour la détermination de la matrice LOP

Pas de temps	Variable	Description
0-58	ev_net_f	EV avant impôt et coût de la marge de solvabilité
59-60	ev_n_av_f	EV après impôt et avant cout de la marge de solvabilité

Tableau 5 : Variables corrélées à la PVFP

Où EV (*Embedded value*) = *ActifNet Réévalué* + *PVFP* – *Coût du capital à immobiliser*

La formule de l'EV montre le lien avec la PVFP et justifie la corrélation entre ces deux variables sur toute la durée de projection.

4.1.3 Calibrage par rapport au BEL et la PVFP

Le calcul des corrélations entre les variables de la base de données des cash flows déterministes et le BEL et la PVFP (voir annexe - Corrélations à la BEL et la PVFP pour le détail des résultats) permet d'obtenir les variables indiquées dans le tableau ci-après pour la détermination de la matrice LOP.

Pas de temps	Variable	Description
0-24	ev_net_f	EV après impôt et après coût de la marge de solvabilité
25-37	nb_dc	Nombre de décès du mois
38-42	nb_cnt_f	Nombre de contrats en cours à la fin de t
43-53	v_act_gest	Valeur actuelle des actes de gestion
54-60	nb_cnt_f	Nombre de contrats en cours à la fin de t

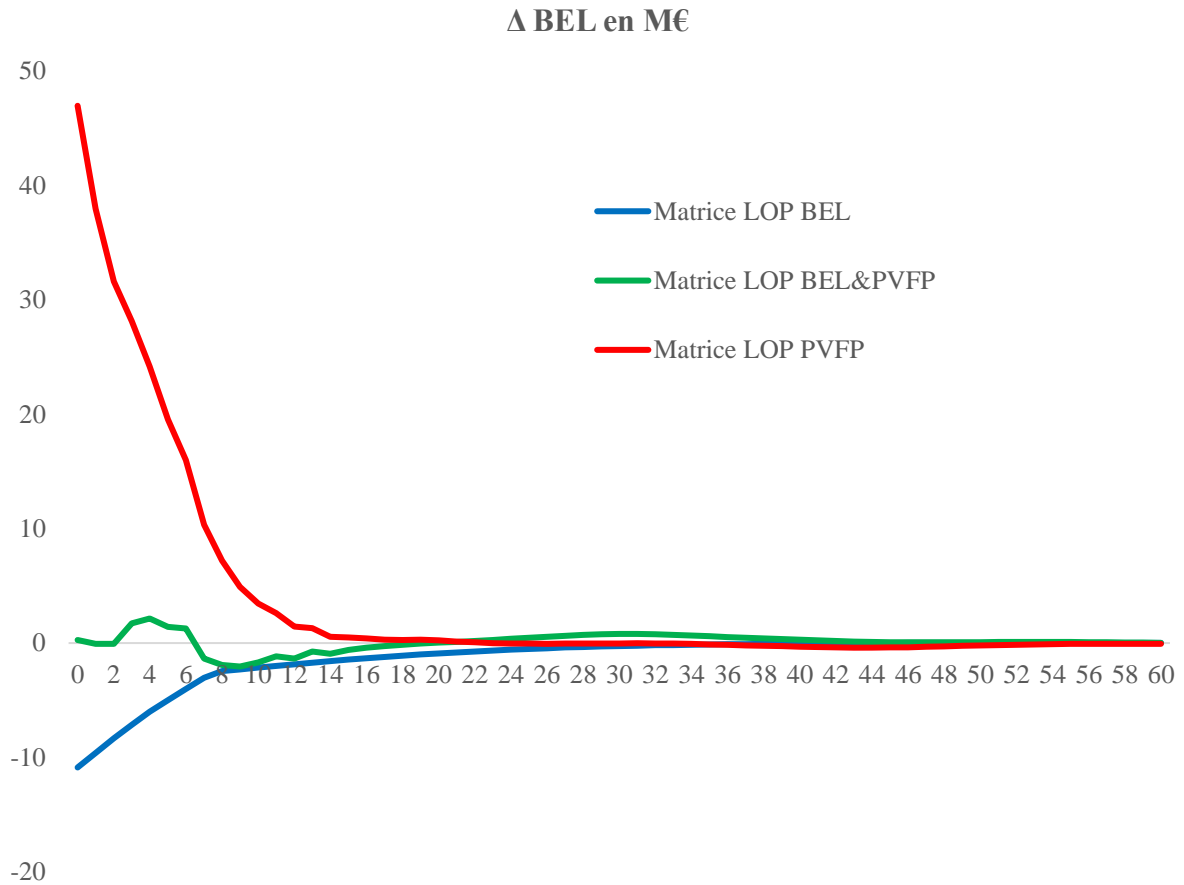
Tableau 6 : Variables corrélées au BEL et la PVFP

4.1.4 Analyse des résultats

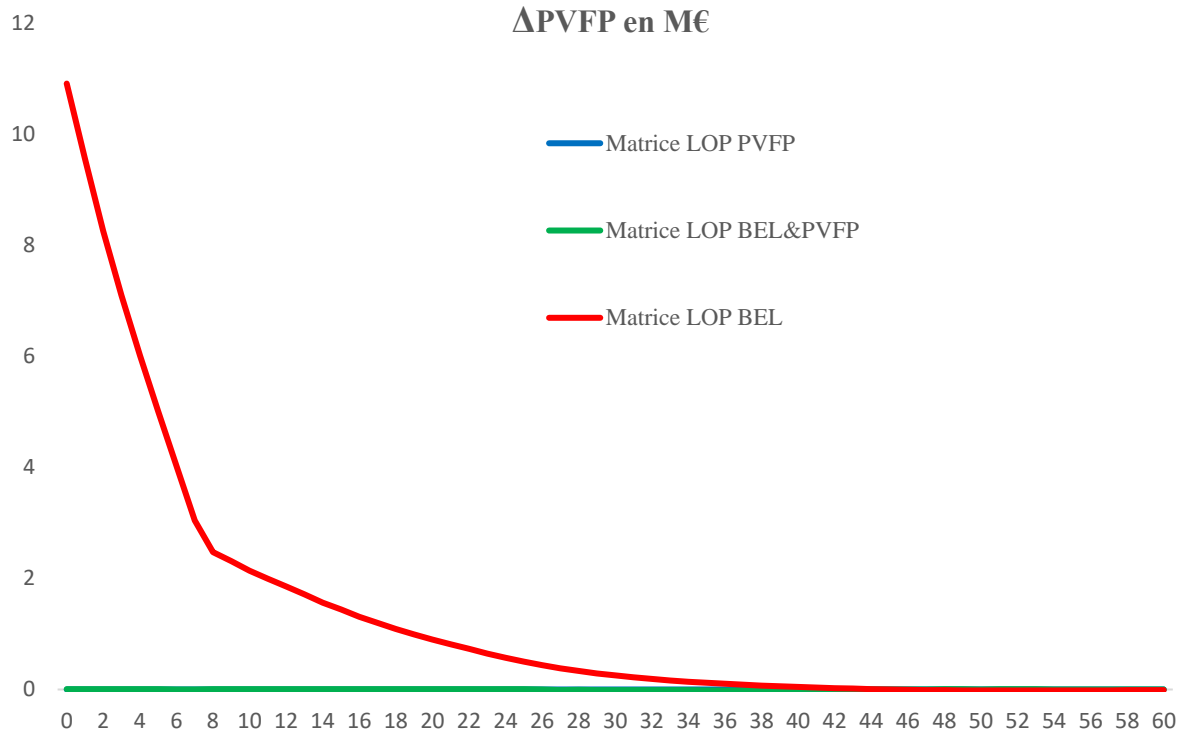
Avec les trois matrices LOP nous allons exécuter l'algorithme LOP et comparer les résultats pour identifier la meilleure matrice LOP, c'est-à-dire celle qui est la plus corrélée au BEL et à la PVFP.

- Analyse sur toute la durée de projection

Les figures ci-dessous donnent l'écart entre le BEL et la PVFP avant agrégation et le BEL et la PVFP après agrégation selon que la matrice LOP est issue de la corrélation avec le BEL, la PVFP ou le BEL et la PVFP sur toutes les années de projection.



Jusqu'à $t = 6$, le BEL basé sur la matrice LOP PVFP s'écarte du BEL avant agrégation. Les autres BEL sont à peu près identiques au BEL avant agrégation sur toute la durée de projection.



Les écarts entre la PVFP calculée avant agrégation et les PVFP après agrégation, calculées grâce aux matrices LOP BEL & PVFP d'une part et la matrice LOP PVFP d'autre part sont nuls sur toute la durée de projection.

L'écart entre la PVFP avant agrégation et la PVFP calculée grâce à la matrice LOP BEL est de plus d'un million d'euros jusqu'à $t=8$. Après cette année, l'écart est inférieur à un million d'euros et devient nul autour de $t=35$.

- Analyse à $t=0$

Les résultats de cette étape du calibrage (à $t=0$) sont regroupés dans le tableau ci-après. ϵ_{Bel} et ϵ_{PVFP} représentent respectivement les écarts relatifs par rapport au BEL et à la PVFP avant agrégation tels que définis au 3.2.2 Processus LOP dans le modèle LSIM-Etape 4

Technique	Critères	BEL (€)	ϵ_{Bel}	PVFP (€)	ϵ_{PVFP}	MP
Données avant agrégation		818 096 469		124 417 575		18 525
Corrélation (Pearson)	BEL	807 222 306	1%	135 333 677	9%	4 015
	PVFP	865 072 070	6%	124 417 720	0%	4 274
	BEL & PVFP	818 370 729	0%	124 421 328	0%	4 230

Tableau 7 : Synthèse des vecteurs LOP par critères et impacts BEL et PVFP

Les meilleurs résultats sont ceux obtenus lorsque les variables utilisées sont celles issues de la corrélation entre le BEL et la PVFP. La qualité des résultats est aussi confirmée quand on effectue les comparaisons à un niveau plus détaillé (Van primes, Van rachat, Van coûts d'acquisition, etc), voir annexe 3 Comparaison des indicateurs constituant le BEL, pour plus de détails. En revanche, avec un taux de compression de 77% (1-4230/18 525), le nombre de MP après agrégation reste élevé. Nous devons trouver d'autres pistes d'agrégation, cette fois basées sur les critères de regroupement des données.

4.2 Calibrage en entrée du modèle.

Dans cette partie, nous allons tester différentes manières de regrouper les données en entrée de l'algorithme LOP. Nous nous servons d'un algorithme de classification non supervisé que nous appliquerons d'une part sur les cash-flows de passif et d'autre part sur les composantes principales des MP.

4.2.1 Pré-Clustering sur les cash-flows de passif

Rappelons d'abord le principe du clustering avant de l'appliquer à notre étude.

Le principe général du partitionnement de données avec le clustering, consiste à maximiser la similarité entre les observations au sein d'un même groupe tout en minimisant les similarités entre différents groupes pour ainsi créer ce qu'on appelle des partitions homogènes. Le processus de partitionnement passe par quelques étapes notables :

- Représentation du jeu de données et sélection d'attributs : il s'agit de la première étape à réaliser. Elle consiste à évaluer le nombre de classes qui partitionnent le jeu de données initial et à sélectionner les attributs et les variables à utiliser dans l'algorithme d'agrégation.
- Choix d'une mesure de similarité ou de proximité : il est nécessaire de définir ce que sont les individus similaires, proches entre eux. Pour cela, une fonction distance entre les objets algébriques est à définir pour l'algorithme d'agrégation.
- Groupement de données : cette étape consiste à affecter des individus à un groupe donné de manière dure (i.e hard clustering), ou d'une manière floue (i.e fuzzy clustering). Cela dépend du type d'algorithme utilisé dans l'agrégation.

4.2.1.1 Critères d'homogénéité des groupes : le théorème de

Huygens

Les méthodes d'agrégation utilisent des critères d'homogénéité. Ces derniers reposent sur la notion de moments d'inertie d'ordre deux pour caractériser les groupes à former selon la dispersion des données.

Soient $X_1, \dots, X_i, \dots, X_n$ les objets classés dans K classes C_1, \dots, C_k (avec $K < n$). En considérant X_G le centre de gravité de ce nuage de points et X_{iG} le centre de gravité de la classe C_i , l'inertie totale est :

$$I = \sum_{i=1}^n d^2(X_i - X_G) = \sum_{i=1}^n \|X_i - X_G\|^2$$

Les classes étant distinctes, l'inertie totale est la somme des inerties de chaque classe :

$$I = \sum_{k=1}^K \sum_{i \in C_k} \|X_i - X_{kG}\|^2 \text{ avec } i \in C_k$$

En s'appuyant sur le théorème de Huygens, on obtient :

$$\begin{aligned} I &= \sum_{k=1}^K \sum_{i \in C_k} \|X_i - X_{kG} + X_{kG} - X_G\|^2 \text{ avec } i \in C_k \\ &= \sum_{k=1}^K \sum_{i \in C_k} (\|X_i - X_{kG}\|^2 + \|X_{kG} - X_G\|^2) \\ &= \sum_{k=1}^K \sum_{i \in C_k} d^2(X_i - X_{kG}) + \sum_{k=1}^K \sum_{i \in C_k} d^2(X_{kG} - X_G) \\ &= I_w + I_b \end{aligned}$$

Avec

$$I_w = \sum_{k=1}^K \sum_{i \in C_k} d^2(X_i - X_{kG})$$

$$I_b = \sum_{k=1}^K \sum d^2(X_{kG} - X_G)$$

I_w mesure l'inertie intra-classe, elle est la somme des distances entre les objets d'une classe et le barycentre de cette classe. Une classe homogène est une classe où l'inertie intra-classe est faible.

I_b mesure l'inertie inter-classes, elle est la somme des distances entre le barycentre des objets et le barycentre de chaque classe. Plus l'inertie inter-classes est grande, plus les classes sont hétérogènes.

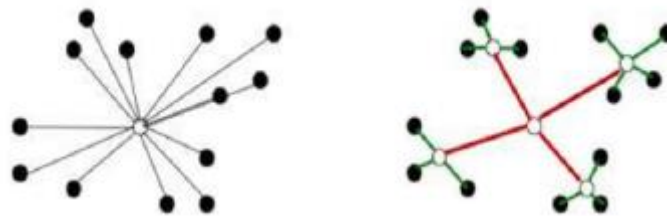


Figure 20 : Inertie intra-classe vs inertie inter-classes

La segmentation en K classes revient à trouver les classes de telle sorte que l'inertie intra-classe soit minimale et l'inertie inter-classes soit maximale. En d'autres termes, la classification consiste à regrouper les individus de sorte à maximiser la similarité intragroupe et à minimiser la similarité inter-groupes.

Il existe plusieurs familles d'algorithmes de partitionnement dont les plus notables sont citées ci-dessous :

- Les algorithmes de clustering **hiérarchique** dont l'objectif est de constituer une hiérarchie de groupes telle que, selon la manière dont on se déplace dans cette hiérarchie, les objets sont plus ou moins similaires. Parmi les algorithmes connus dans cette famille nous pouvons citer les arborescences hiérarchiques ascendante ou descendante.
- Les algorithmes de clustering par **partition** dont l'objectif est de partitionner ou couper l'espace des objets selon un ou plusieurs critères déterminés, pour ainsi former des groupes homogènes y répondant. Il existe dans le cadre de ces algorithmes plusieurs sous-familles comme :
 - L'algorithme des K -means ou des K -médoides,

- Les algorithmes basés sur la densité,
- Les algorithmes basés sur les graphes connexes,
- Les algorithmes basés sur les réseaux de neurones.

Dans cette partie théorique, nous nous contenterons d'expliquer uniquement la méthode de partitionnement K-MEANS. Notre choix d'utiliser cette méthode au-delà des résultats plus probants qu'elle nous donne, s'appuie sur les travaux présentés dans l'article de Mingoti Sueli A. et Lima Joab O [2005], qui exposent une étude comparative de plusieurs méthodes.

4.2.1.2 K-MEANS

4.2.1.2.1 Principe

La méthode de partitionnement des K-Moyennes (K-MEANS, MacQueen, 1967) est l'une des plus répandues et des plus simples à mettre en œuvre. L'objectif est de partitionner les données étudiées notées en K groupes disjoints que l'on améliore itérativement jusqu'à la convergence du critère choisi. La figure suivante illustre ce processus.

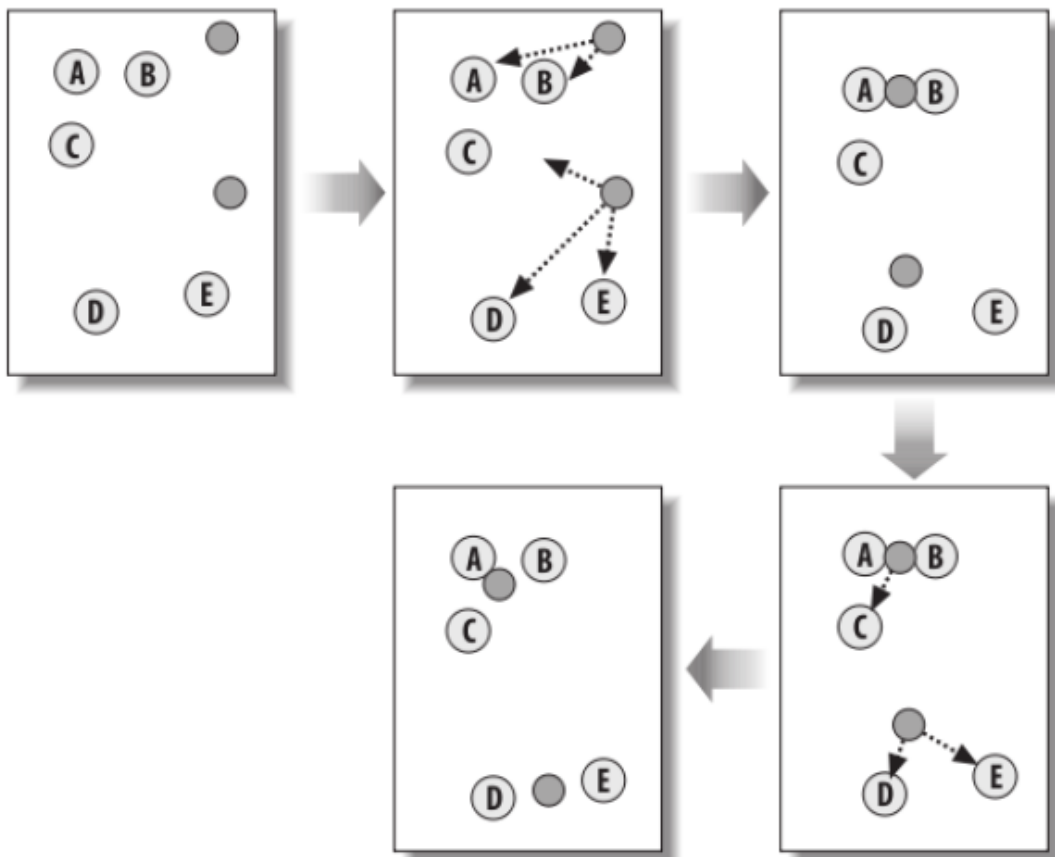


Figure 21 : Classification K-Means

4.2.1.2.2 L'algorithme

L'algorithme classique des K-MEANS peut être résumé en quelques étapes :

- Choisir un nombre de groupes K : cette première étape requiert une analyse préalable car c'est une entrée de l'algorithme,
- Sélectionner aléatoirement $k = 1 \dots K$ objets de la base de données étudiée en les considérant comme centres initiaux (i.e. moyennes des groupes unitaires) : c'est une étape perfectible dans plusieurs autres versions de l'algorithme des K-MEANS,
- Affecter chaque observation à son plus proche centre, en utilisant la distance euclidienne entre chaque observation,
- Pour chaque groupe k , mettre à jour les centres en recalculant les moyennes de chaque groupe après nouvelle affectation.
- Minimiser itérativement l'inertie intra-classe en répétant les étapes 3 et 4 jusqu'à convergence de l'inertie (ou atteinte du nombre maximum d'itérations donné en entrée).

La convergence de l'algorithme est assurée quelle que soit l'initialisation. Une esquisse de démonstration de cela découle de la décroissance de la suite minorée des inerties intra-classes calculées, le minimum étant atteint pour le barycentre des points donc pour leur moyenne

4.2.1.3 Mise en œuvre de l'algorithme du K-Means

L'implémentation du K-Means passe par les étapes suivantes :

- Calcul des cash-flows déterministes de passif pour chaque ligne de MP (18 525).
- Lancement d'un run qui calcule différents indicateurs pour chaque ligne de *model points* (18 525). Ce calcul diffère de celui effectué actuellement en ce que les MP ne sont pas regroupés ici. En effet, dans le calcul actuel, les MP sont regroupés en 185 groupes selon la variable groupe de contrats. Le regroupement étant l'objet du clustering, nous n'effectuons donc pas de regroupement dans un premier temps.
- Choix des indicateurs de classification des modèles
 - Nous avons plusieurs indicateurs en passif seul. 3 indicateurs semblent pertinents, parce qu'ils sont représentatifs du passif d'un bilan d'une compagnie d'assurance vie : le BEL, la PM et la PVFP.
 - BEL

Nous avons présenté le BEL au chapitre 1 (1.1.1.1 Présentation générale de la norme). Nous renvoyons à cette partie pour la description de cet indicateur.

- PM

Les compagnies d'assurance constituent des provisions mathématiques afin de faire face aux engagements pris envers leurs assurés. Le mode de constitution de ces provisions est réglementé par le Code des Assurances (article R331-1 et suivants).

En assurance vie, ces provisions sont calculées en tenant compte des probabilités de décès ou de survie des assurés. Elles sont égales à la différence entre la valeur actuelle probable (VAP) des engagements de l'assureur (VAP_A) et la valeur actuelle probable des engagements pris par l'assuré (VAP_a). D'où la relation fondamentale :

$$- \quad PM(t) = VAP_A(t) - VAP_a(t)$$

- PVFP

Nous avons présenté la PVFP au 4.1.2 Calibrage par rapport à la PVFP . Nous renvoyons à cette partie pour la description de cet indicateur.

- Identification du nombre de groupes à constituer : la technique du « coude »

La méthode du « coude » est la technique la plus connue pour déterminer le nombre de clusters idéal. Elle s'appuie sur la notion d'inertie. De manière pratique on construit un graphique qui en fonction du nombre de clusters associe la distance au carré (Y) entre chaque point et le centroïde dans un groupe. Au fur et à mesure que le nombre de clusters augmente, Y diminue. Quand on analyse le graphique, nous pouvons voir que le graphique va changer rapidement à un moment donné et adopter une forme de coude. A ce niveau on a le nombre optimal de clusters à créer.

Nous avons appliqué cette technique sur nos trois indicateurs et les combinaisons possibles de ces indicateurs. Ce qui donne un total de sept options : BEL, PM, PVFP, BEL&PM, BEL & PVFP, PM & PVFP, BEL&PM& PVFP. Dans les sept cas, le nombre optimal de clusters est quinze (15). Ci-après la représentation de la courbe du nombre de clusters en fonction du coût du modèle.

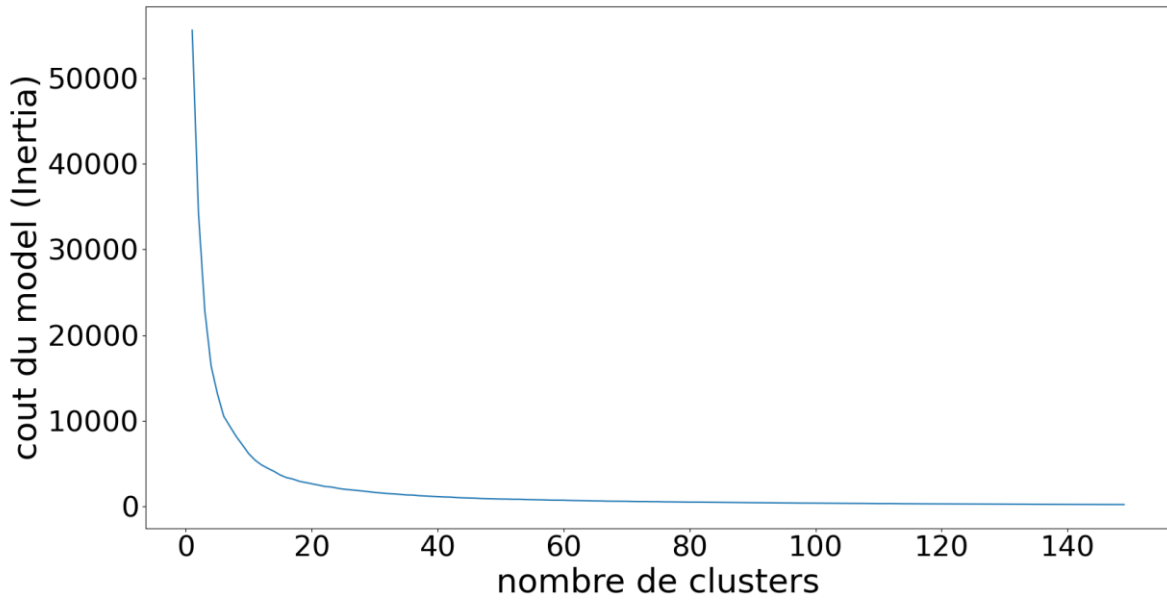


Figure 22 : Nombre de clusters en fonction du coût du modèle (BEL, PM et PVFP)

La présentation des données en trois dimensions permet de mieux visualiser les différents clusters. Les nuages de points étant quasi identiques dans les sept cas, nous présentons ici uniquement le cas du regroupement selon le critère du BEL.

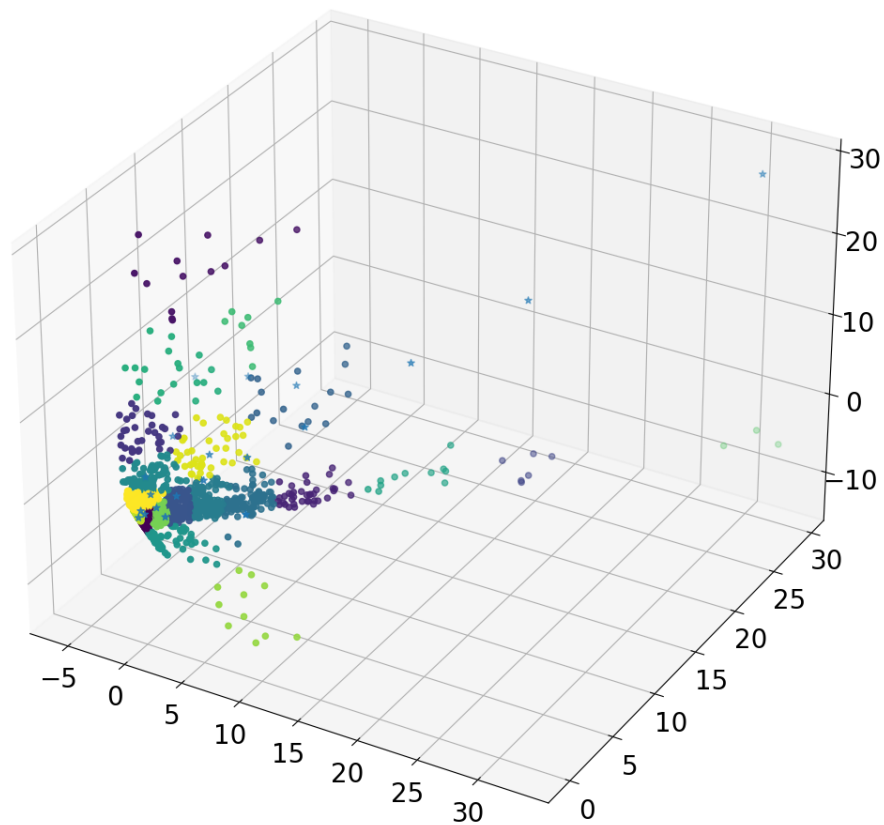


Figure 23 : Nuage de points en trois dimensions

- Présentation des clusters

Le tableau ci-dessous donne les résultats obtenus après lancement de l'algorithme du K-Means.

Cluster	BEL	PM	PVFP	BEL &PVFP	PVFP &PM	BEL &PM	BEL &PM& PVFP
Cluster 1	12 649	11 188	16 307	15 837	13 025	13 678	15 085
Cluster 2	3 143	3 518	825	1 517	2 974	2 754	1 998
Cluster 3	1 235	1 623	574	381	1 139	1 047	576
Cluster 4	575	897	288	311	501	441	261
Cluster 5	296	475	152	136	263	248	259
Cluster 6	213	286	126	93	201	112	114
Cluster 7	120	172	77	71	131	80	80
Cluster 8	98	123	46	42	95	65	39
Cluster 9	84	82	45	42	84	31	31
Cluster 10	40	80	29	33	40	21	28
Cluster 11	28	40	22	23	28	18	17
Cluster 12	24	21	11	14	24	11	12
Cluster 13	11	11	10	11	11	10	12
Cluster 14	6	6	8	11	6	6	10
Cluster 15	3	3	5	3	3	3	3
Total	18 525	18 525	18 525	18 525	18 525	18 525	18 525

Tableau 8 : Clusters basés sur les variables run MP / MP

Sur ce tableau, on peut faire les observations suivantes :

- La majorité des données sont concentrées dans le premier cluster (plus de 60% des données) dans tous les cas.
- Les 5 derniers clusters (cluster 15, cluster 14, cluster 13, cluster 12 et cluster 11) ont quasiment la même taille dans les 7 cas. On observe toutefois une exception lorsque l'indicateur de lancement de l'algorithme est la PM. Dans ce dernier cas, le cluster 11 a deux fois plus de données que dans les autres cas.

Nous allons exécuter l'algorithme LOP dans chacun de ces 7 cas. Les variables LOP en input étant celles que nous avons obtenues à la première étape du calibrage (4.1.4 Analyse des résultats). Le tableau ci-après présente les résultats obtenus.

Technique	Critères	BEL (€)	ε Bel	PVFP (€)	ε PVFP	MP
Données avant agrégation		818 096 469		124 417 575		18 525
Clustering (K-Means)	BEL	579 978 307	-29%	124 335 320	-0,1%	576
	PM	639 839 004	-22%	124 407 947	0,0%	610
	PVFP	539 311 370	-34%	124 376 114	0,0%	492
	BEL &PVFP	519 329 738	-37%	124 230 556	-0,2%	473
	PVFP &PM	512 349 252	-37%	124 273 607	-0,1%	574
	BEL &PM	737 028 768	-10%	124 403 306	0,0%	500
	BEL &PM& PVFP	892 832 540	9%	124 543 663	0,1%	494

Tableau 9 : Agrégation sur la base du Clustering K-Means

Sur ce tableau, nous pouvons faire les constats suivants :

- L'écart relatif par rapport à la PVFP reste acceptable dans tous les cas.
- Avec le clustering, nous arrivons à réduire considérablement le nombre de MP par rapport aux résultats obtenus dans le *Tableau 7 : Synthèse des vecteurs LOP par critères et impacts BEL et PVFP*.
- Cependant l'écart relatif par rapport au BEL se dégrade. Ceci probablement à cause du cluster 1, comme nous l'avons indiqué, qui représente plus de 60% des données.

Pour améliorer les résultats du clustering, il faudra envisager une technique pour mieux répartir les données. Nous allons effectuer pour cela le clustering non plus sur les données en sortie du run passif seul MP par MP, mais plutôt sur les MP en entrée de l'étape 1 (voir Figure 18 : Processus LOP). Les MP possédant 13 variables, il faudra au préalable effectuer une analyse en composantes principales pour déterminer les variables sur lesquelles le clustering sera effectué.

4.2.2 Pré-Clustering sur les composantes principales des MP

4.2.2.1 Lien entre l'ACP et l'algorithme du K-Means

L'analyse en composantes principales (ACP) permet de réduire le nombre de variables (V) d'une base de données tout en préservant la variance (voir Annexe 4 : Principe général de l'ACP). L'algorithme du K-Means permet quant à lui de réduire le nombre de « points » (P) en résumant plusieurs points en fonction de leur distance à la moyenne (voir 4.2.1.2. K-MEANS). Donc si le jeu de données comporte (P) points avec (V) caractéristiques, l'ACP vise à

compresser les caractéristiques (V) alors que le clustering vise à compresser les points (P). A priori ces deux techniques semblent donc ne pas être liées. Cependant, il est courant d'appliquer l'ACP pour réduire le nombre de dimensions avant d'appliquer le K-Means (Zha et al., 2002).

Ding & He (2004)⁶ dans un article, énoncent le théorème suivant (voir démonstration dans l'article, théorème 3.3) concernant le lien entre l'ACP et le K-Means :

le sous-espace formé par les centroïdes de la solution du clusters K-Means est incorporé dans les premières K- 1 composantes principales.

$$\mathbf{S}_d = \sum_{k=1}^{K-1} \lambda_k \mathbf{u}_k \mathbf{u}_k^T$$

où $\vec{\mathbf{u}}_k$ sont les axes principaux de variabilité du nuage λ_k les valeurs propres et \mathbf{S}_d le sous-espace formé par les centroïdes.

Ce théorème implique que la réduction des dimensions au travers l'ACP trouve automatiquement le meilleur sous-espace des centres des clusters du K-Means. L'ACP peut être utilisée pour initialiser les itérations de K-Means. Ceci explique pourquoi il est bénéfique de faire une ACP avant un K-Means.

Ils concluent cet article avec la proposition suivante : les composantes principales sont les solutions continues aux indicateurs d'appartenance à un cluster discret pour le clustering K-Means.

En outre sur la base d'un échantillon de données, ils observent que :

- moins il y a de composantes principales, plus le K-Means est précis.
- le K-Means après ACP sur les données centrées semble donner des meilleurs résultats que sur les données non centrées.

4.2.2.2 Mise en œuvre du K-Means sur les composantes principales

Les principes de l'ACP sont rappelés en annexe (Annexe 4 : Principe général de l'ACP), nous allons présenter ici directement son implémentation dans notre cas. Pour cela, nous allons étudier les corrélations entre les variables pour justifier la pertinence d'effectuer une ACP,

⁶ Ding C & He X [2004], *Op cit.*

choisir le type d'ACP à effectuer et le nombre de composantes principales à retenir. Ceci nous permettra d'identifier les axes sur lesquels nous effectuerons notre analyse.

4.2.2.2.1 Matrice de corrélation entre variables

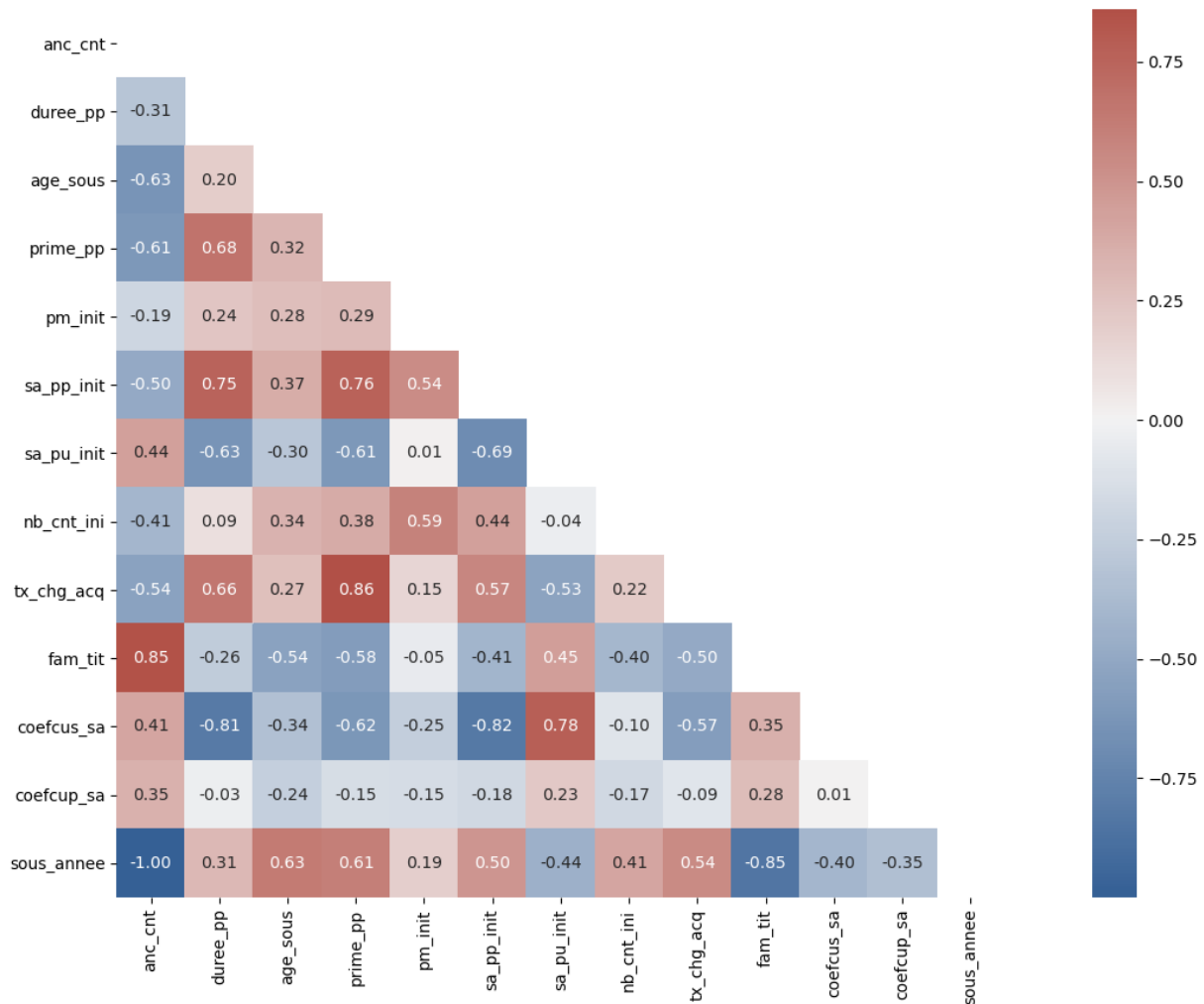


Tableau 10 : Matrice de corrélation

On peut remarquer que :

- L'ancienneté et l'année de souscription sont parfaitement corrélées.
- Le taux de charge d'acquisition et les primes périodiques sont fortement corrélés (0.86). Il en est de même entre l'ancienneté et le taux d'intérêt technique (0.85), et le montant des capitaux investis en prime périodique et la durée de versement des primes périodiques (0.75).

Puisque les variables ne sont pas toutes corrélées entre elles et que les coefficients de corrélation ne sont pas tous proches de zéro, nous avons intérêt à réduire les dimensions de notre portefeuille grâce à une ACP.

4.2.2.2.2 Type d'ACP

Nous allons faire une ACP normée afin d'éviter qu'une variable prédomine sur les autres. Nous nous appuyerons sur quelques sorties de Python afin d'interpréter ces résultats et de choisir les composantes principales qui serviront à notre étude.

4.2.2.2.3 Nombre de composantes principales à retenir

Diverses techniques sont utilisées pour définir le nombre de composantes principales qui seront retenues. Une première approche consiste à choisir le nombre minimum de composantes qui explique le maximum d'information. Le tableau ci-dessous donne les valeurs propres avec le pourcentage de variances expliquées.

	Valeurs propres	Variances expliquées	Variances expliquées cumulées
Composante 1	4,36	34%	34%
Composante 2	2,74	21%	55%
Composante 3	1,58	12%	67%
Composante 4	1,06	8%	75%
Composante 5	0,76	6%	81%
Composante 6	0,65	5%	86%
Composante 7	0,5	4%	90%
Composante 8	0,45	3%	93%
Composante 9	0,38	3%	96%
Composante 10	0,29	2%	98%
Composante 11	0,18	1%	99%
Composante 12	0,07	1%	100%
Composante 13	0	0%	100%
Total	13	100%	

Tableau 11 : Valeurs propres et variances expliquées

En examinant donc la proportion expliquée par chaque composante, nous constatons que le premier plan (les deux premières composantes) ne porte que 55 %. Cependant, les quatre premières composantes portent trois quarts de l'information (75%). Cela nous conduit à retenir ces quatre premières composantes.

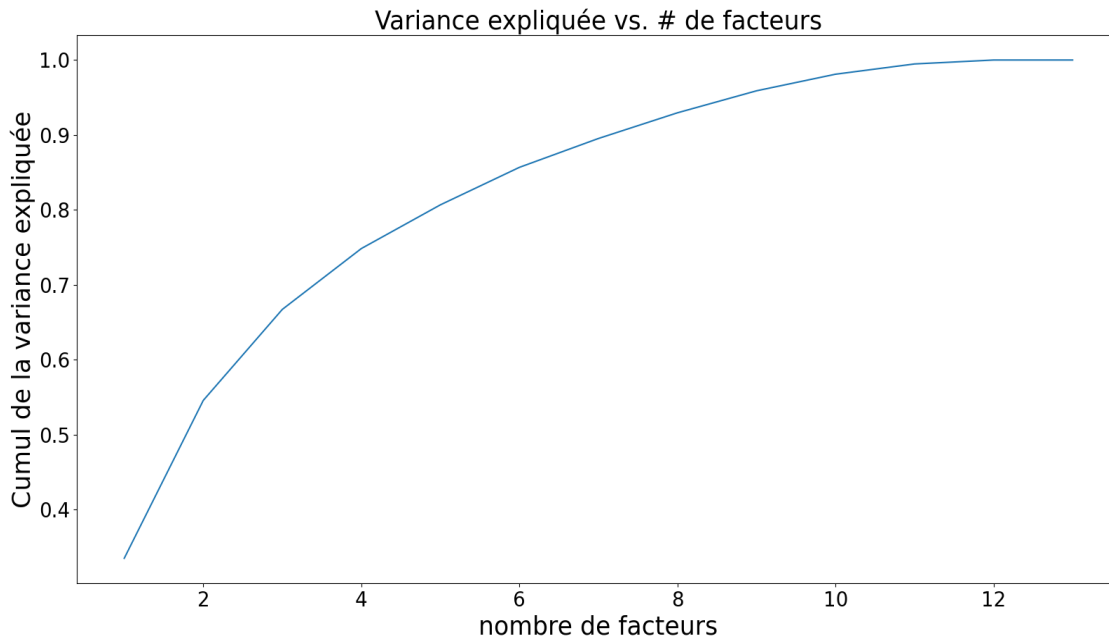


Figure 24 : Nombre de facteurs en fonction de la variance expliquée

Une autre approche pour le nombre de composantes est l'approche de Kaiser. Cette approche se fonde sur l'analyse des valeurs propres : dans une ACP normée, la somme des valeurs propres étant égale au nombre de variables (13 dans notre cas), leur moyenne vaut 1. Un axe est intéressant si sa valeur propre est supérieure à 1. En d'autres termes, un axe est intéressant s'il contribue plus qu'une variable prise individuellement. Le critère de Kaiser confirme le choix de 4 composantes effectué précédemment.

Ce choix est enfin confirmé par la méthode des « coudes », comme le montre le graphique ci-après. On peut observer plusieurs coudes sur le graphique. Parmi ces coudes on retrouve un coude avec quatre facteurs.

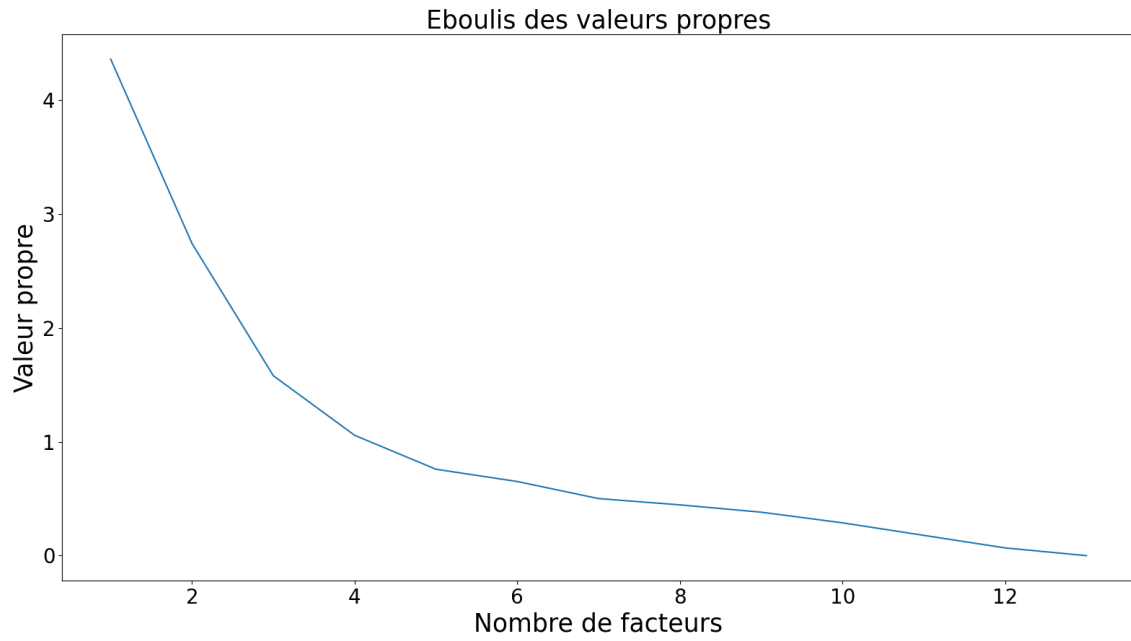


Figure 25 : Eboulis des valeurs propres

4.2.2.2.4 Représentations des variables

Représentons dans un premier temps les variables sur le plan constitué des deux premières composantes.

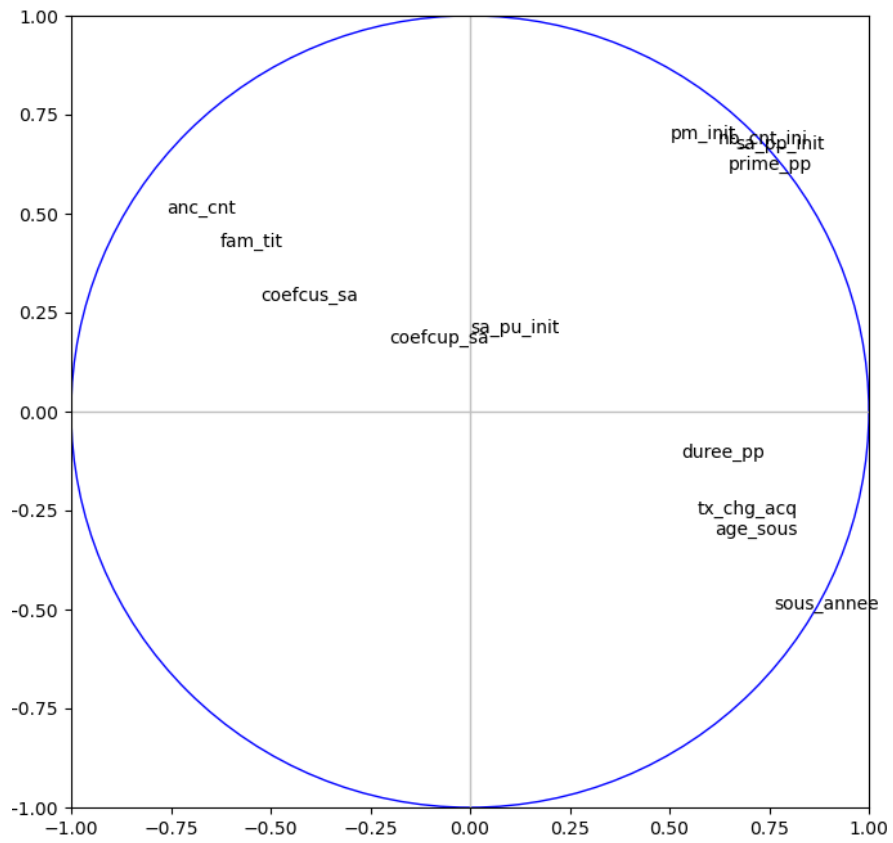


Figure 26 : Cercle de corrélation axes 1 et 2

Interprétation des composantes :

- Première composante

Le tableau ci-dessous donne la contribution des différentes variables à cette composante.

Variables	CONTR_1
anc_cnt	13%
sous_annee	13%
sa_pp_init	10%
prime_pp	10%
fam_tit	9%
nb_cnt_ini	9%
age_sous	9%
tx_chg_acq	7%
duree_pp	6%
coefcus_sa	6%
pm_init	6%
coefcup_sa	1%
sa_pu_init	0%

Tableau 12 : Représentation des variables sur le premier axe factoriel

On remarque que les principales variables qui contribuent à cette composante sont les variables « anc_cnt », « sous_annee », « sa_pp_init » et « age_sous ». Par ailleurs on remarque que ces variables sont bien représentées sur le cercle de corrélation. En effet, les variables « sous_annee » et « sa_pp_init » sont quasiment sur le cercle de corrélation, les variables « anc_cnt » et « age_sous » ne sont pas très loin du cercle de corrélation. En outre l'observation des individus qui contribuent le plus à cette composante montre qu'ils ont au moins deux valeurs importantes sur les variables « anc_cnt » et « age_sous ». On peut donc conclure que la durée des engagements caractérise cette première composante.

- Diagramme à moustache « sous_annee ».

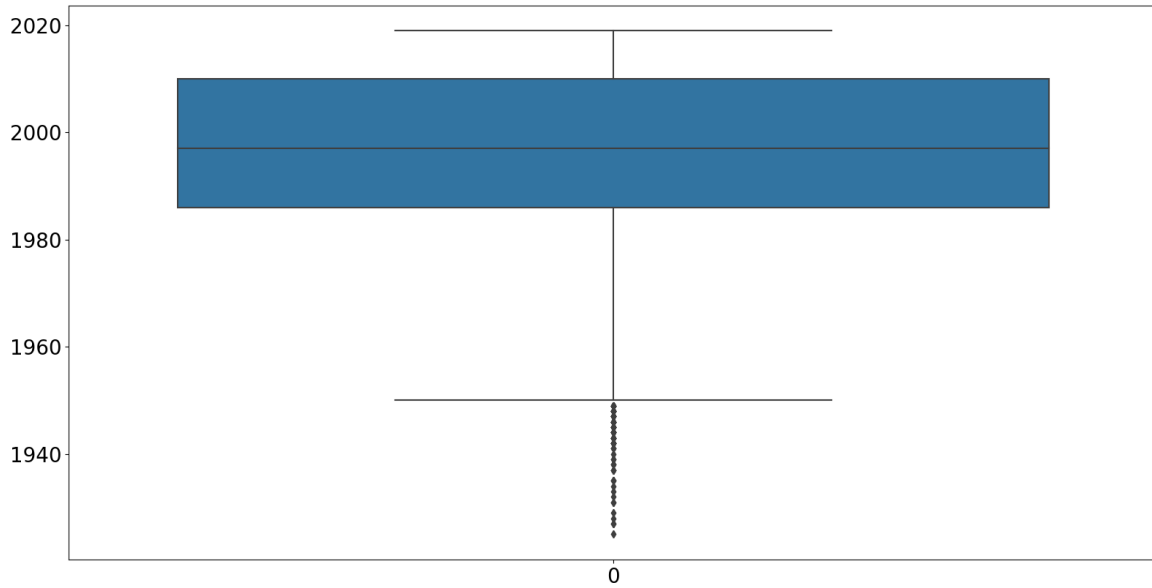


Figure 27 : Diagramme à moustache de la variable « sous_annee » (année de souscription)

- Diagramme à moustache « sa_pp_init »

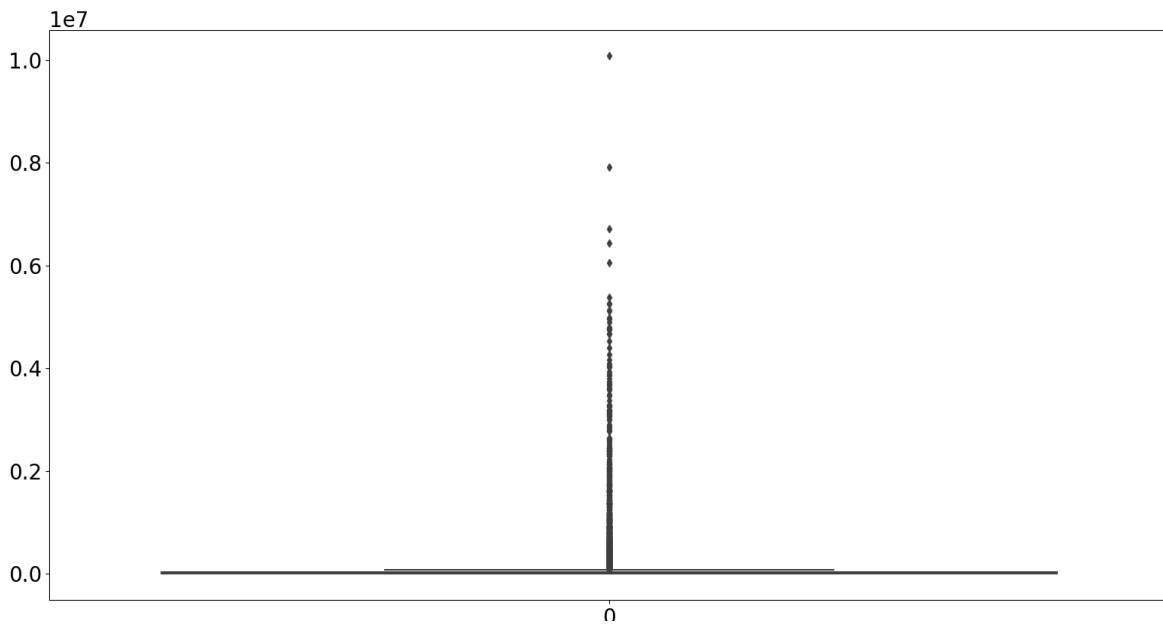


Figure 28 : Diagramme à moustache de la variable «sa_pp_init »

L'analyse des diagrammes à moustache montre que la variable « sous_annee » est moins dispersée pour segmenter les *model points*. Nous choisissons donc cette variable pour regrouper

nos MP. Après ce regroupement, nous allons lancer l'algorithme LOP avec les variables retenues les plus corrélées au BEL et à la PVFP (cf. 4.1.3).

- Deuxième composante

Les principales variables qui contribuent à cette composante sont : « pm_init », « nb_cnt_ini », « sa_pp_init ». Ces variables sont également bien représentées sur le plan factoriel.

Variables	CONTR_2
pm_init	17%
nb_cnt_ini	17%
sa_pp_init	16%
prime_pp	14%
anc_cnt	9%
sous_annee	9%
fam_tit	6%
age_sous	4%
coefcus_sa	3%
tx_chg_acq	2%
sa_pu_init	1%
coefcup_sa	1%
duree_pp	0%

Tableau 13 : Représentation des variables sur le second axe factoriel

Pour les mêmes raisons que précédemment, nous retenons la variable « pm_init » pour cet axe.

Représentons maintenant les variables sur le plan constitué de la troisième et la quatrième composantes.

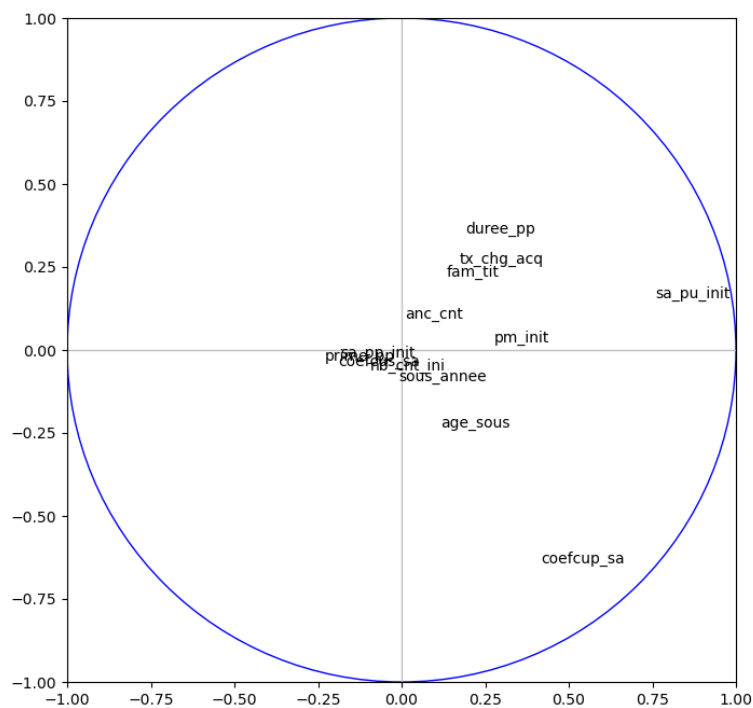


Figure 29 : Cercle de corrélation axes 3 et 4

- Troisième composante

La principale variable qui contribue à cette composante est la variable « coefcup_sa ».

Variables	CONTR_3
coefcup_sa	19%
sa_pu_init	18%
duree_pp	18%
coefcus_sa	18%
tx_chg_acq	9%
sous_annee	5%
anc_cnt	5%
age_sous	3%
pm_init	3%
fam_tit	1%
prime_pp	0%
sa_pp_init	0%
nb_cnt_ini	0%

Tableau 14 : Représentation des variables sur le troisième axe factoriel

- Quatrième composante

La principale variable qui contribue à cette composante est la variable « sa_pu_init »

Variables	CONTR_4
sa_pu_init	54%
coefcup_sa	17%
pm_init	7%
prime_pp	5%
duree_pp	4%
coefcus_sa	3%
sa_pp_init	3%
tx_chg_acq	3%
fam_tit	2%
age_sous	1%
nb_cnt_ini	1%
sous_annee	0%
anc_cnt	0%

Tableau 15 : Représentation des variables sur le quatrième axe factoriel

Ayant identifié les principales variables représentant les quatre premières composantes principales, nous pouvons maintenant effectuer le clustering sur la base de ces variables.

4.2.2.2.5 Cluster sur la base de l'ACP

Comme précédemment, le nombre optimal de clusters est de 15 comme indiqué dans la figure ci-après.

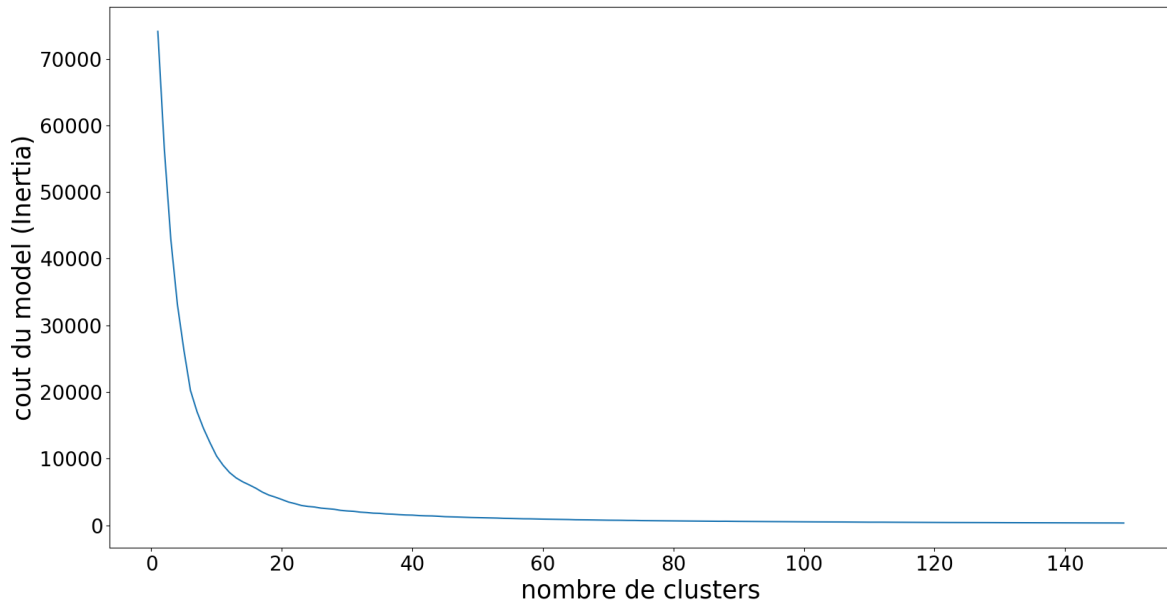


Figure 30 : Nombre de clusters- K-Means sur ACP

Le tableau suivant donne la répartition des données dans les différents clusters.

Cluster	Première composante	2 premières composantes	3 premières composantes	4 premières composantes
Cluster 1	2 018	3 630	4 472	16 686
Cluster 2	2 005	3 255	4 059	757
Cluster 3	1 999	3 167	3 156	218
Cluster 4	1 716	2 700	2 830	155
Cluster 5	1 703	2 063	983	142
Cluster 6	1 689	974	843	133
Cluster 7	1 606	710	611	131
Cluster 8	1 421	639	573	131
Cluster 9	1 296	612	467	83
Cluster 10	1 026	383	334	44
Cluster 11	563	218	126	22
Cluster 12	531	108	50	15
Cluster 13	442	49	17	5
Cluster 14	392	14	3	2
Cluster 15	118	3	1	1
Total	18 525	18 525	18 525	18 525

Tableau 16 : Clusters basés sur ACP sur les MP

On peut effectuer les observations suivantes :

- Le cluster 1 comporte moins de 25% des données dans tous les cas, sauf quand le regroupement est effectué sur la base des 4 premières composantes.
- Les données dans le cas d'un regroupement sur la base de la première composante principale sont très uniformes et peu dispersées : plus on ajoute des composantes, moins la répartition des données est uniforme et plus les données sont dispersées entre les clusters. Nos données permettent ainsi de confirmer l'observation faite par Ding & He (2004)⁷.
- Les dix premiers clusters contiennent entre 2 018 et 1 026 données. Le dernier cluster regroupe plus de 100 individus

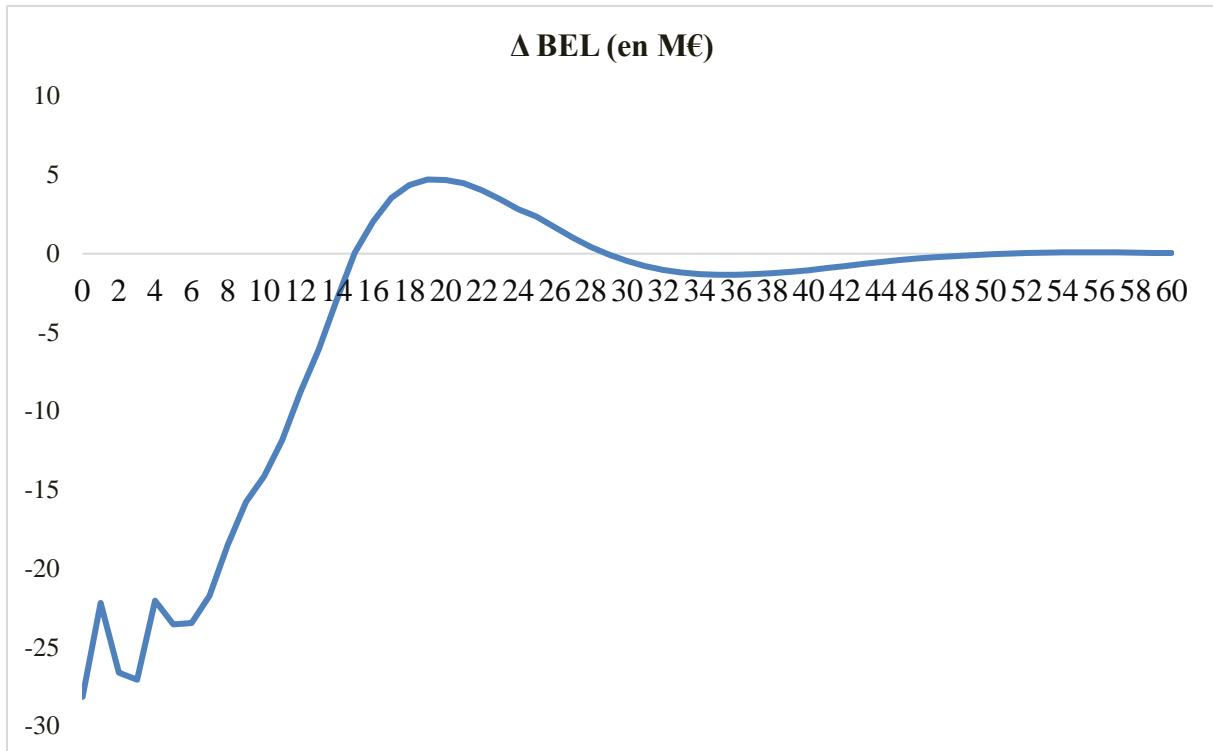
Nous allons maintenant exécuter l'algorithme LOP sur ces données. Nous obtenons les résultats suivants à t=0 :

Technique	Critères	BEL (€)	ε Bel	PVFP (€)	ε PVFP	MP
Données avant agrégation		818 096 469		124 417 575		18 525
Clustering (K-Means) + ACP	Première composante	789 955 751	-3%	124 400 301	0%	903
	2 premières composantes	663 749 376	-19%	124 331 376	0%	738
	3 premières composantes	624 027 512	-24%	124 359 304	0%	691
	4 premières composantes	417 909 577	-49%	124 269 096	0%	617

Tableau 17 : Agrégation sur la base du Clustering K-Means après ACP

Le résultat obtenu après regroupement sur la base de la première composante donne le meilleur compromis erreur BEL, erreur PVFP et compression des MP. Qualité de résultat confirmée par le graphique suivant sur les 60 autres pas de temps, qui compare la BEL après agrégation et la BEL avant agrégation.

⁷ Ding C & He X [2004], *Op cit.*



Pour apprécier la qualité d'agrégation sur toute la durée de projection, nous utiliserons un indicateur, $\delta_{Indicateur}$, défini comme suit :

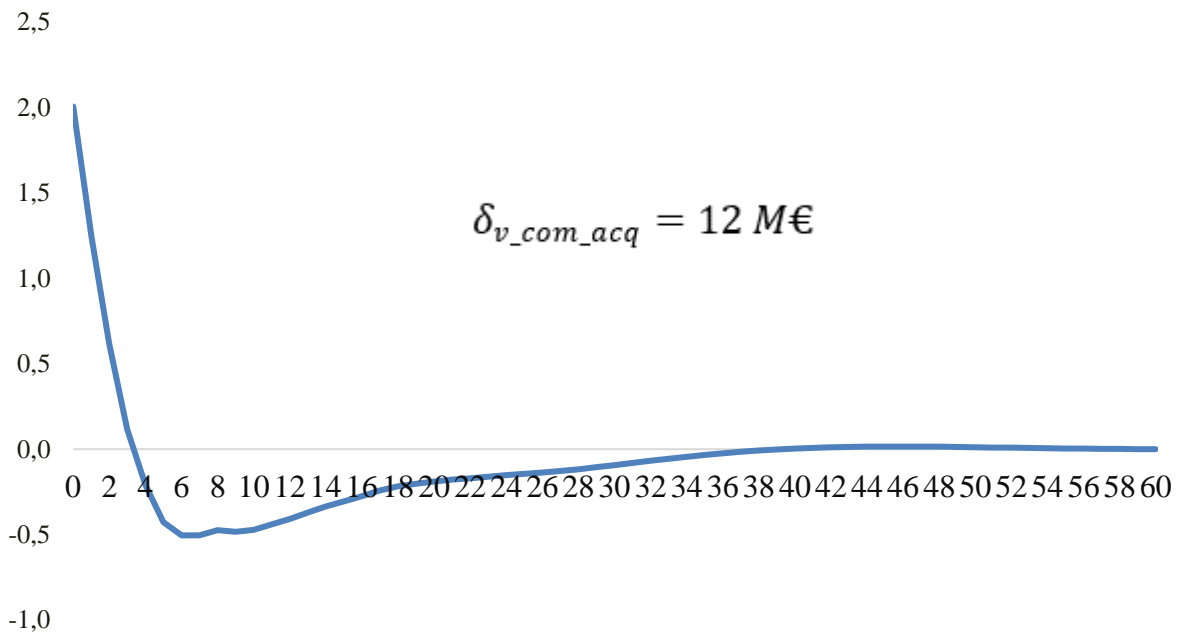
$$\delta_{Indicateur} = \sum_{t=0}^{60} |Indicateur_{(après) t} - Indicateur_{(avant) t}|$$

Pour le BEL, on a :

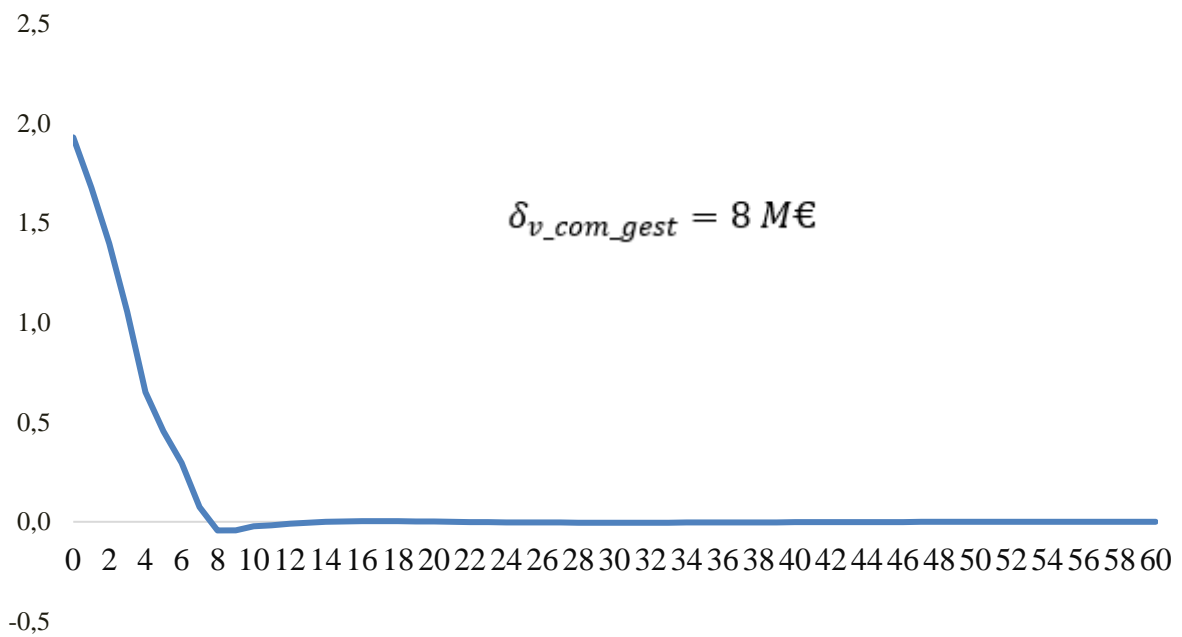
$$\delta_{BEL} = \sum_{t=0}^{60} |BEL_{(après) t} - BEL_{(avant) t}| = 330 \text{ M€}$$

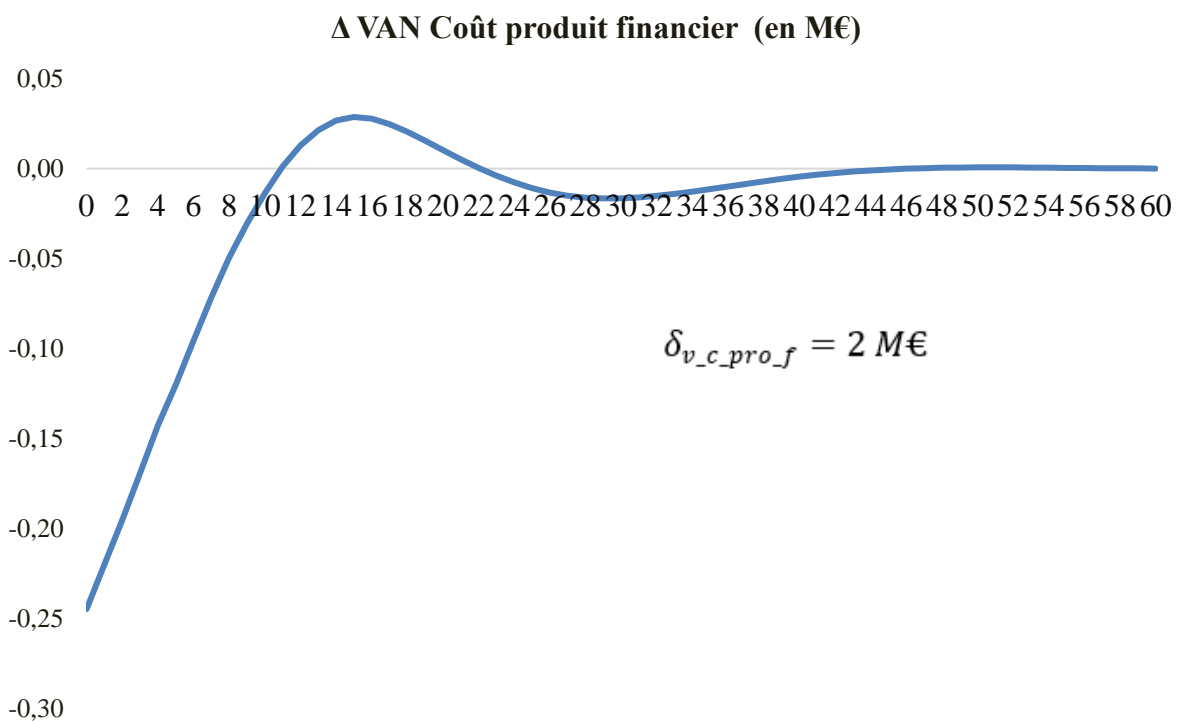
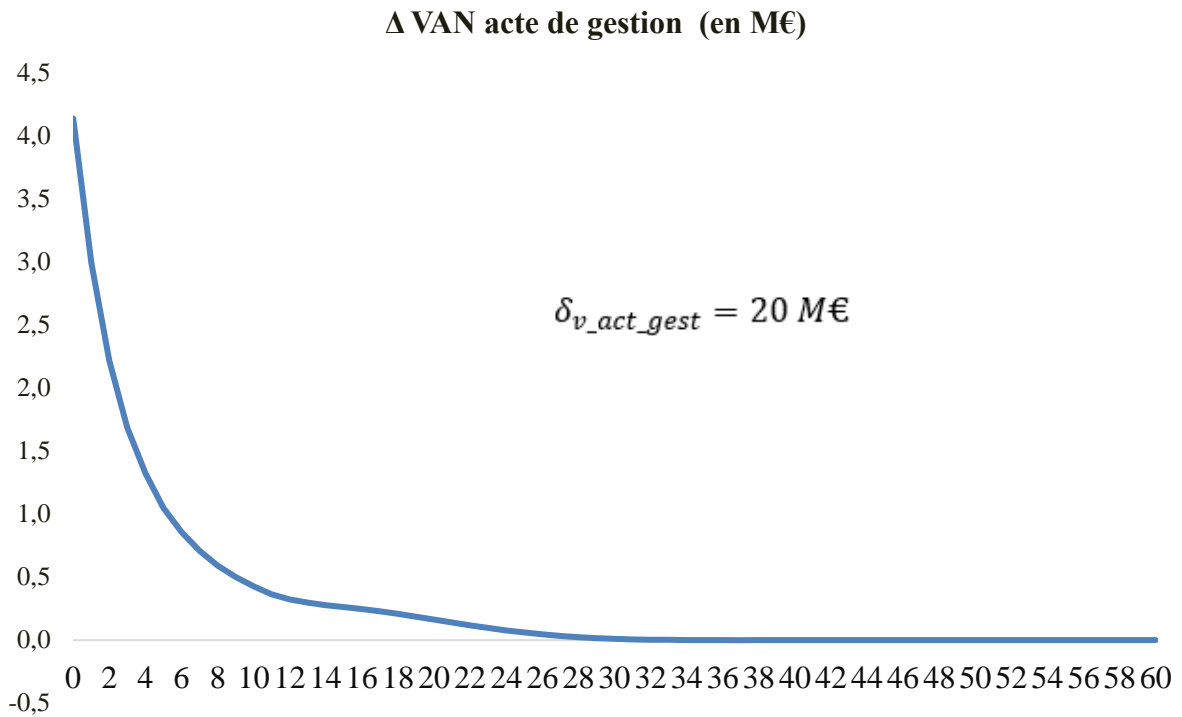
La comparaison des résultats peut être aussi faite au niveau des différentes variables qui constituent le BEL. Les graphiques suivants montrent que sur ces différentes variables, l'erreur entre les données avant agrégation et les données après agrégation est faible sur les 60 pas de temps.

Δ VAN Coût d'acquisition (en M€)

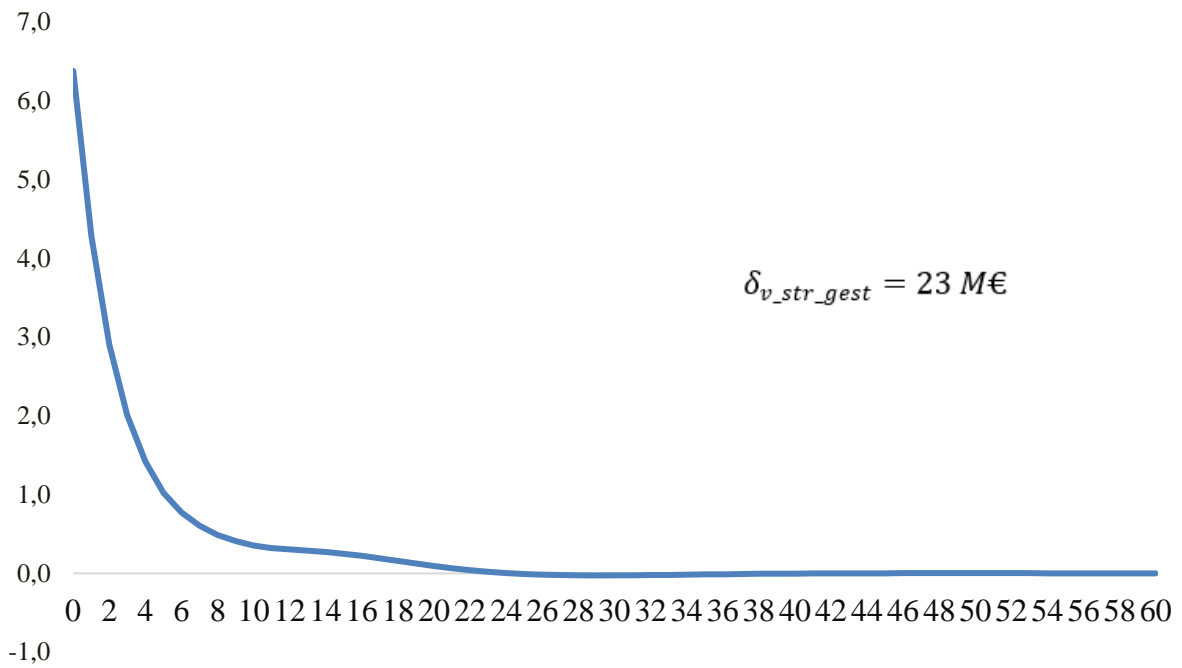


Δ VAN Coût de gestion (en M€)

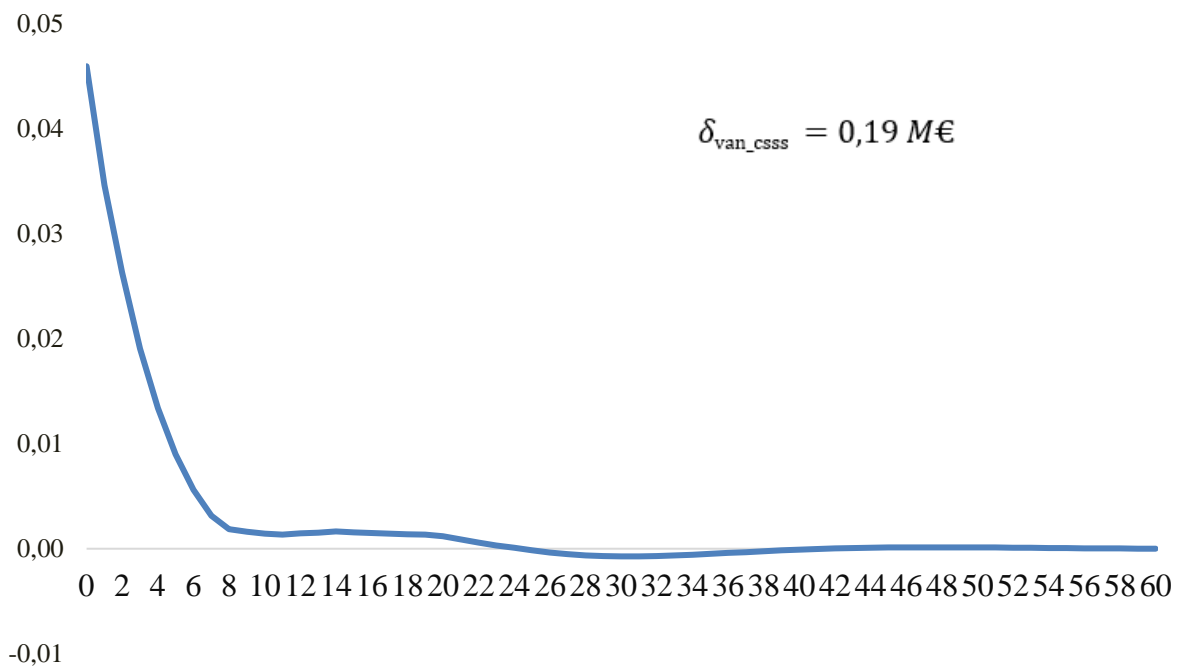




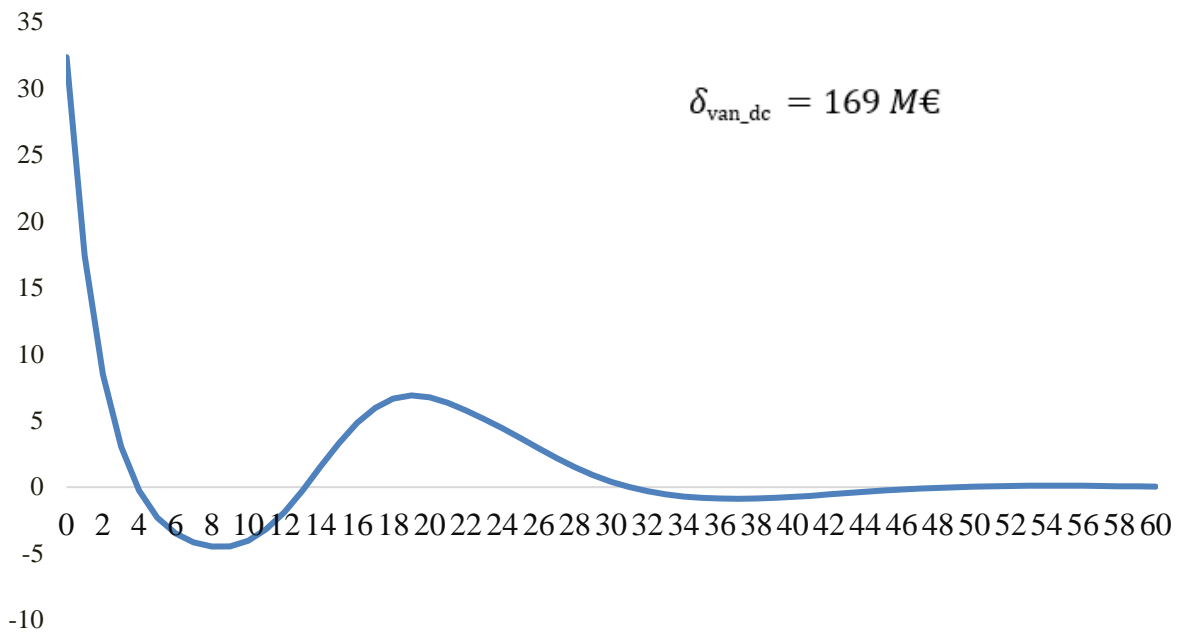
Δ VAN coût structure de gestion (en M€)



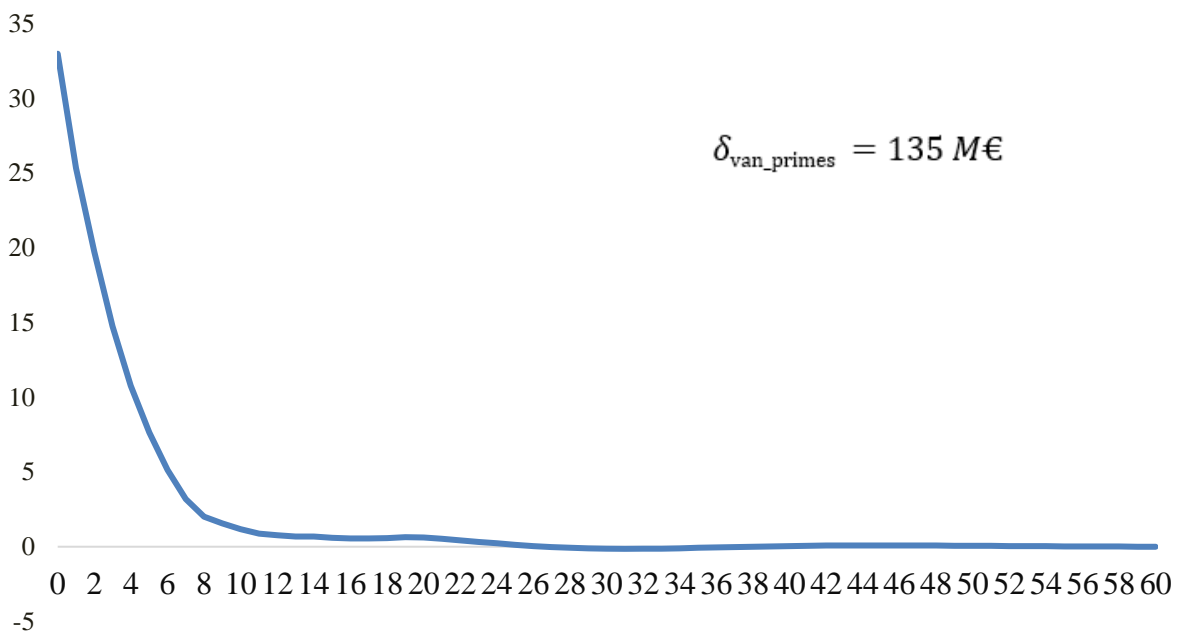
Δ van csss (en M€)



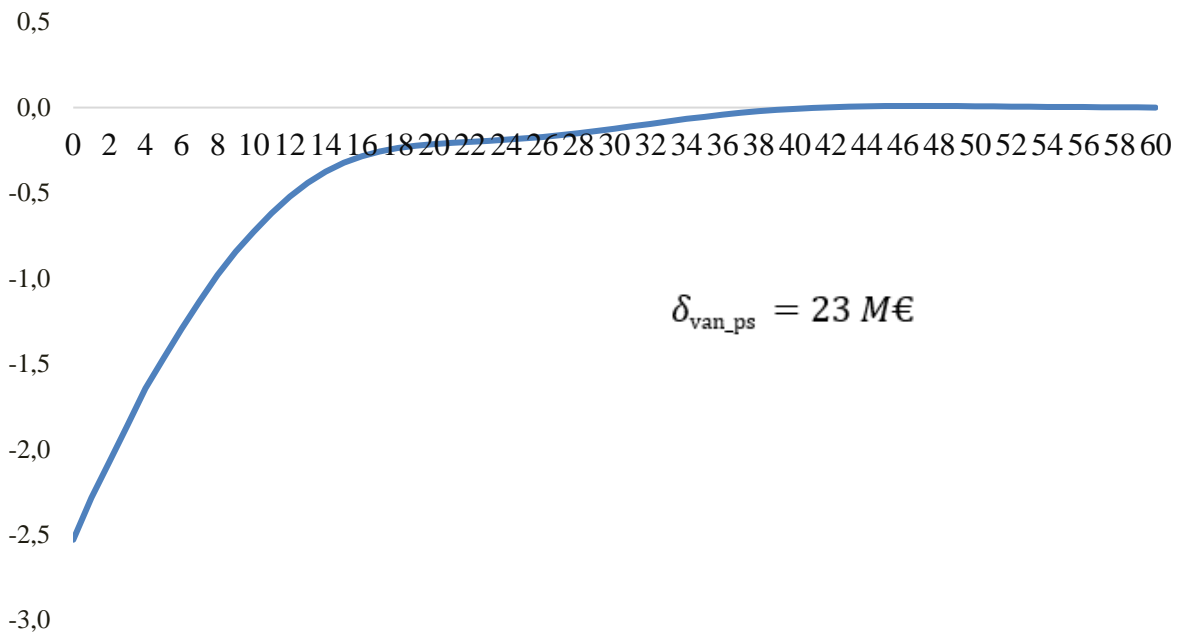
Δ van dc (en M€)



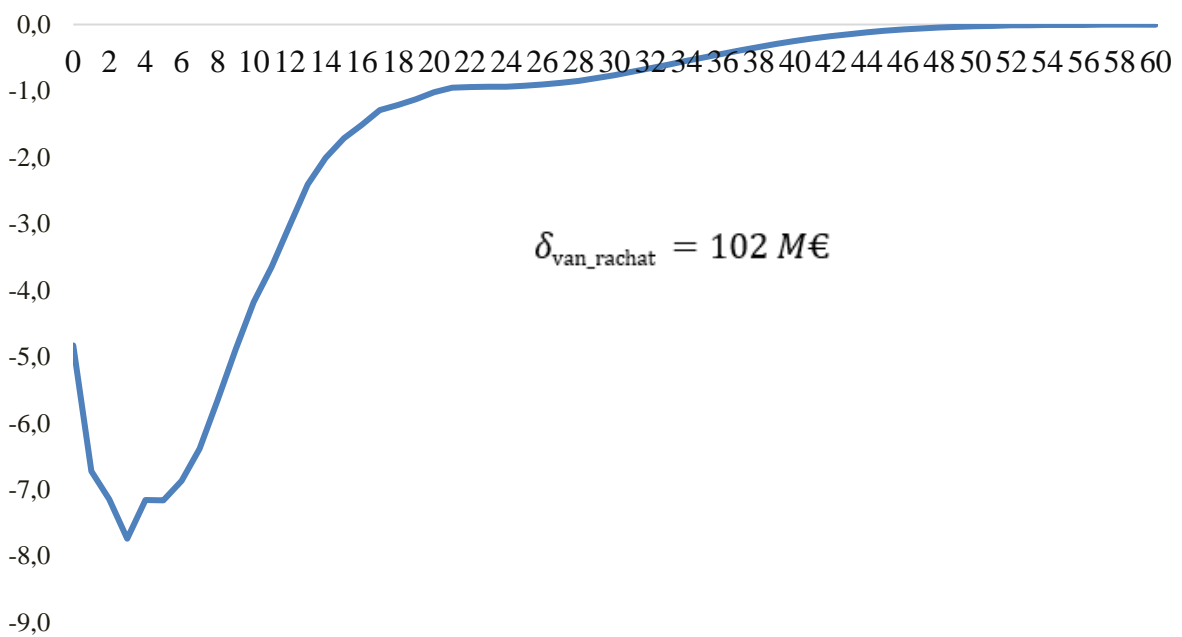
Δ van primes (en M€)

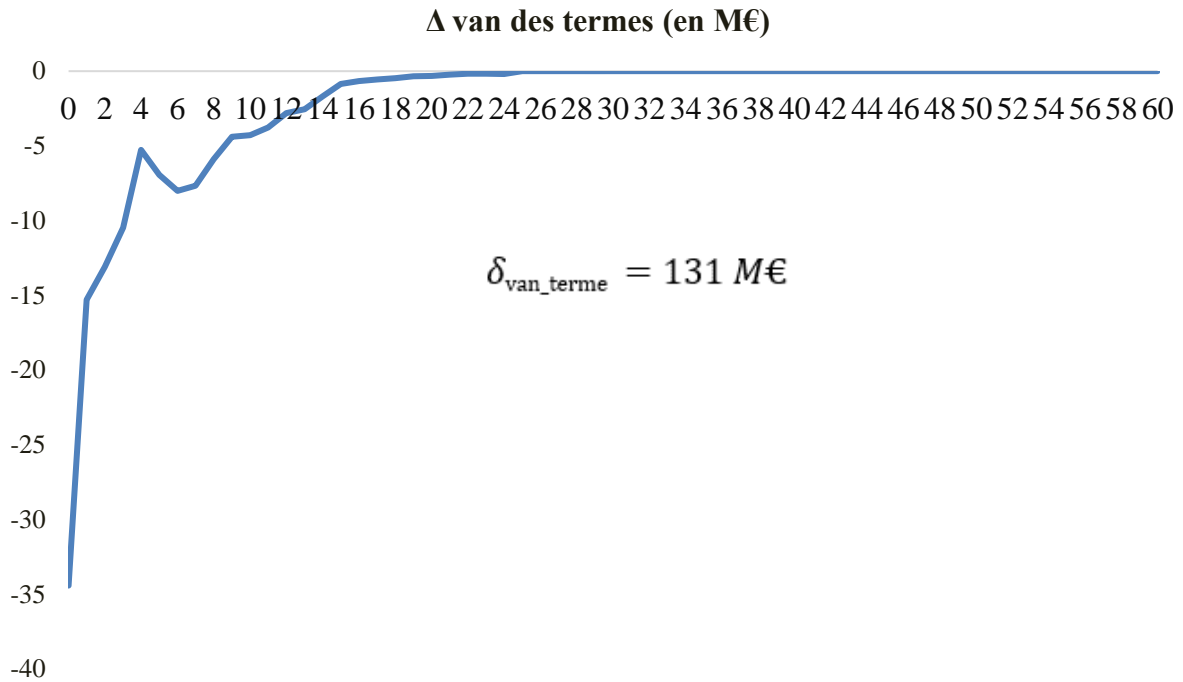


Δ van prélèvements sociaux (en M€)



Δ van des rachats (en M€)





4.2.3 Analyse des résultats

En appliquant l'algorithme de K-Means sur les cash flows déterministes nous avons pu obtenir moins de 650 MP après compression dans toutes les configurations, soit un taux de compression d'environ 97%. Même si l'erreur relative par rapport à la PVFP est acceptable ($|\epsilon| < 1\%$), l'erreur relative par rapport à la BEL s'est fortement dégradée avec cette approche.

L'application du K-Means sur la première composante des MP nous permet de moins dégrader le BEL (erreur relative = -3%), tout en ayant une erreur liée à la PVFP nulle (0%) et d'avoir une bonne agrégation des données, 903 MP, soit un taux de compression de 95%. Nous proposons de retenir cette configuration pour le calibrage de notre modèle.

CONCLUSION

L'agrégation des MP tient une place importante dans les travaux de clôture IFRS 17. Elle permet de simplifier le processus actuariel, d'avoir une modélisation plus fine des passifs et d'obtenir des temps de calcul raisonnables. Obtenir des temps de calcul raisonnables nécessite de calibrer les paramètres du modèle d'agrégation. C'est ce que nous avons effectué dans le cadre de ce mémoire sur le périmètre prévoyance individuelle.

Nous avons effectué le calibrage en deux étapes : premièrement en choisissant les variables de calcul de l'algorithme et deuxièmement, en identifiant la meilleure façon de les regrouper (Pré-Clustering).

Nous avons trouvé des variables de calcul de l'algorithme corrélées à la fois au BEL et la PVFP. Ces variables nous ont permis d'obtenir un taux de compression de 77% et des écarts relatifs par rapport à la BEL et la PVFP presque nuls. Ce niveau d'agrégation n'étant pas optimal, nous avons identifié d'autres leviers de calibrage au travers des critères de regroupement des MP.

Le regroupement des données grâce au K-Means sur la base des critères BEL & PVFP & PM a permis d'obtenir un taux de compression des MP de 97%, avec un écart relatif d'environ 0% pour la PVFP, mais de 9% pour le BEL. Le regroupement des données grâce au K-Means sur la première composante principale des MP a permis d'obtenir la meilleure agrégation : Taux de compression MP : 95%, écart relatif BEL 3% et écart relatif PVFP 0%. Même si le taux de compression est plus faible dans ce dernier cas (95% contre 97%), la prise en compte de l'erreur BEL (3% contre 9%) nous amène à retenir ce dernier résultat pour le calibrage de notre modèle.

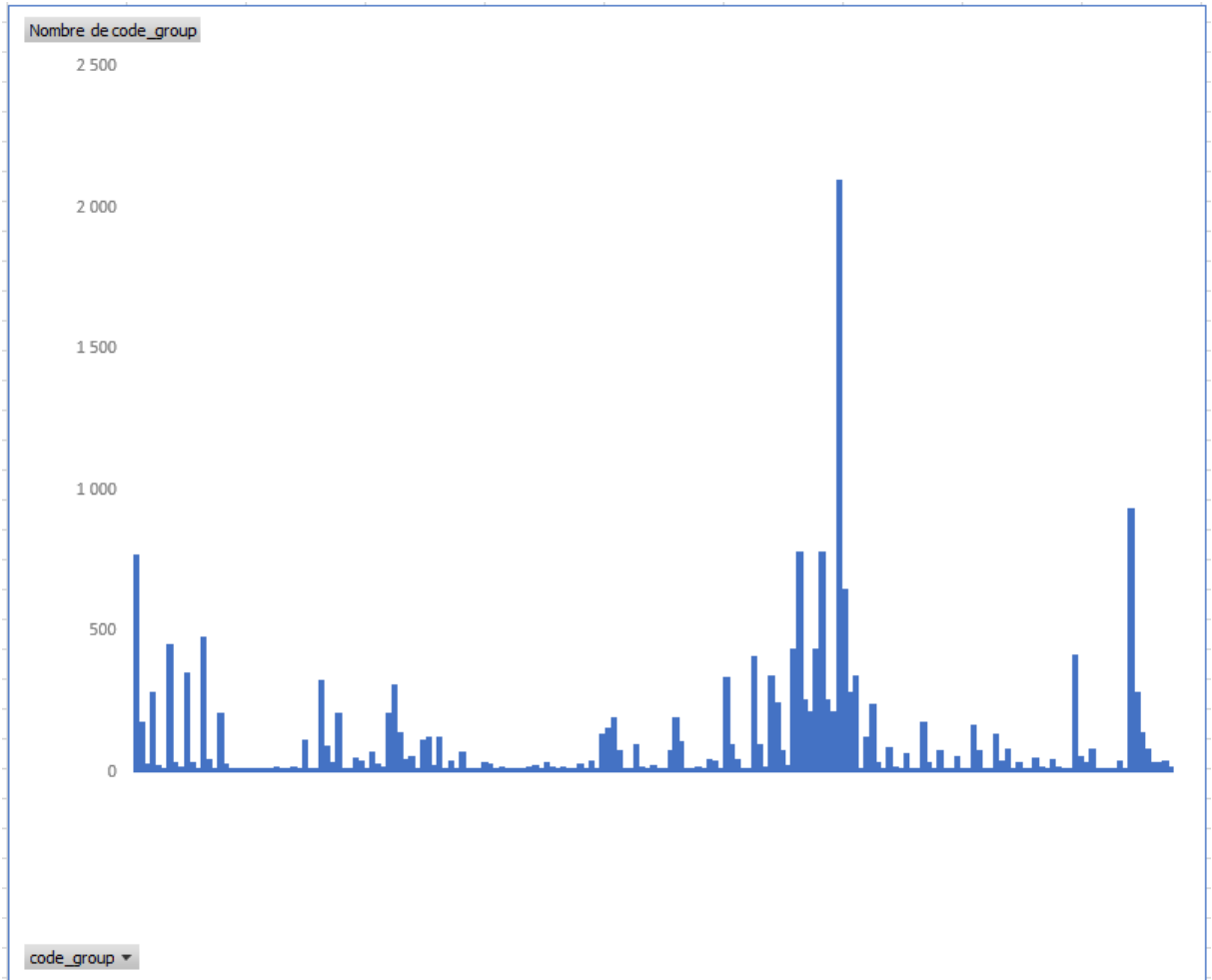
Notre étude permet de montrer ainsi qu'en combinant le LOP avec un algorithme de clustering on arrive à des meilleurs résultats qu'en utilisant uniquement le LOP. Par ailleurs, en effectuant le clustering sur les MP on obtient des meilleurs résultats qu'en les effectuant sur les variables de cash-flows déterministes basées sur ces MP. Une application de la démarche de calibration proposée dans cette étude sur d'autres périmètres permettra de généraliser ces conclusions. Notons aussi que notre étude a été limitée à une analyse des sorties déterministes. Nous n'avons pas effectué une analyse stochastique, notamment des temps de run stochastiques. Mais a priori plus le nombre de MP est réduit, plus le temps de run sera court, toutes choses étant égales par ailleurs.

Pour finir, les résultats de cette étude suscitent une question sur la constitution même des MP. En effet plutôt que remplacer l'algorithme LOP, ne faudrait-t-il pas plutôt repenser le processus de création des MP pour respecter les contraintes d'agrégation qu'impose un modèle *full* ALM ? Dans ce cas, le pre-clustering s'effectuerait sur la base des contrats plutôt que sur les MP.

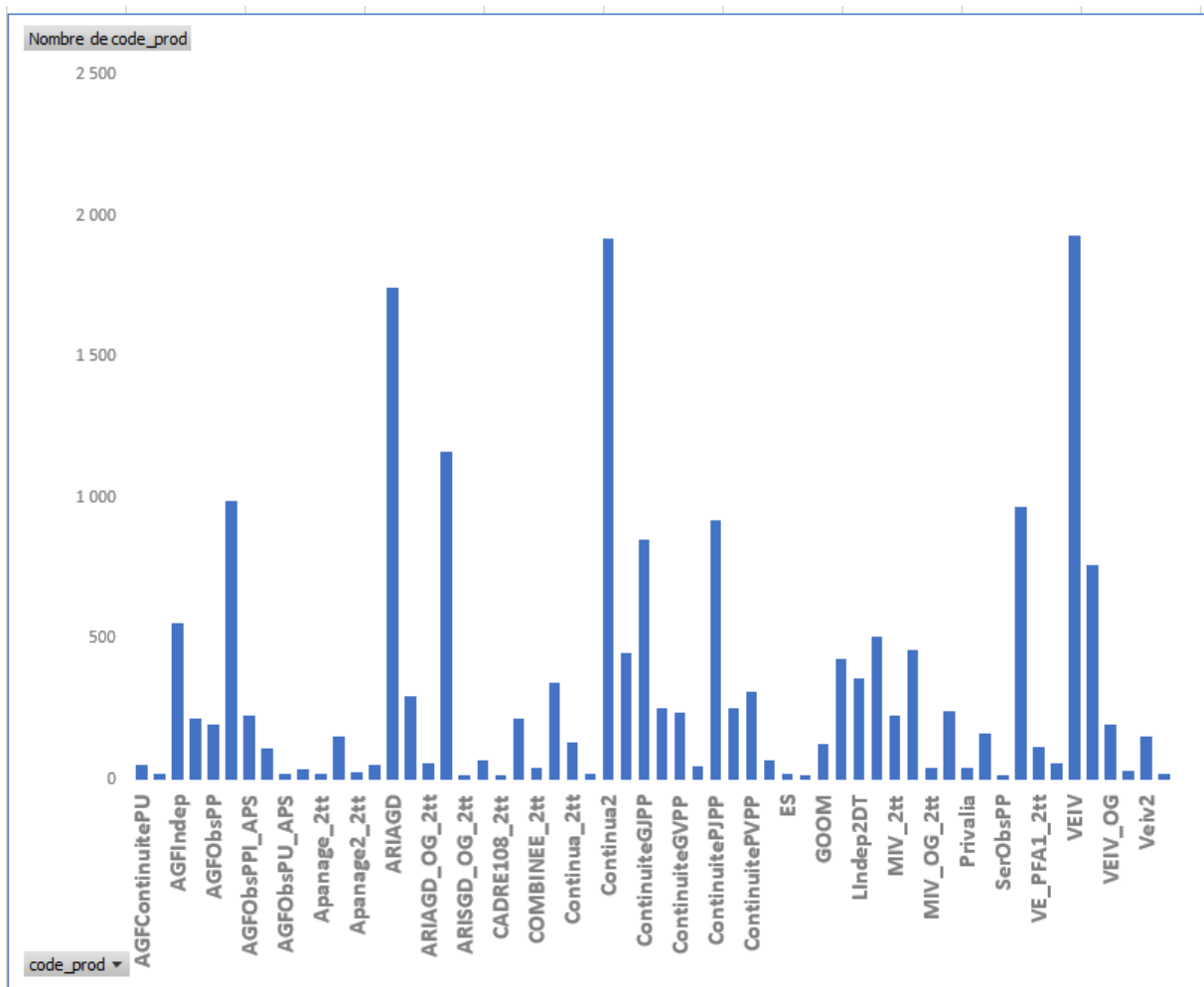
ANNEXES

Annexe 1 : Statistiques descriptives

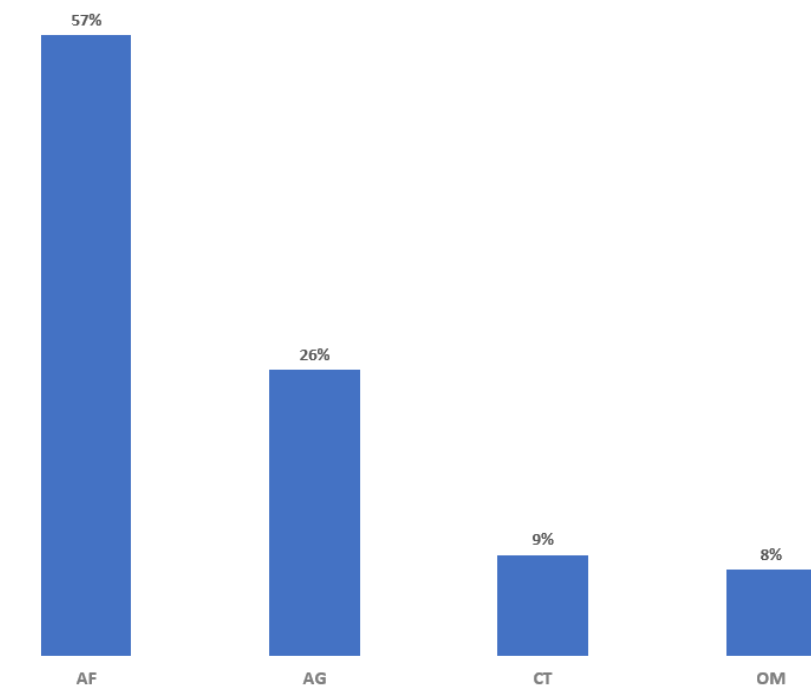
- Variables qualitatives
- Code groupe



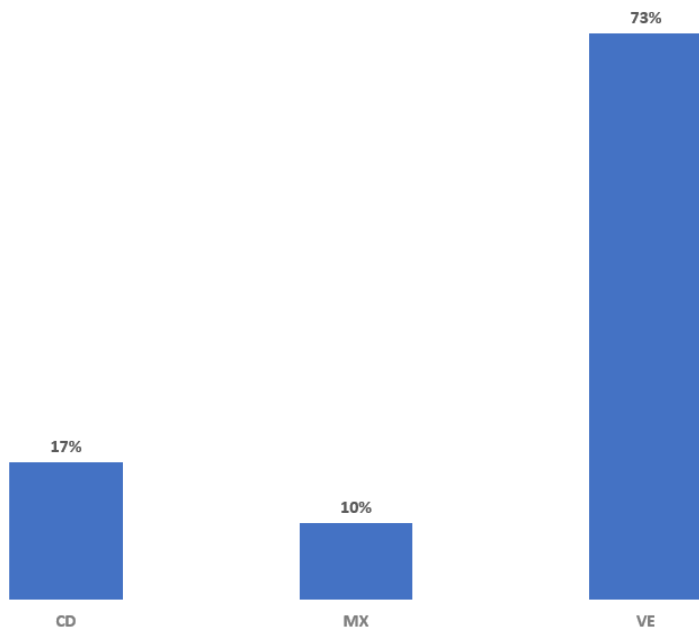
- Code produit



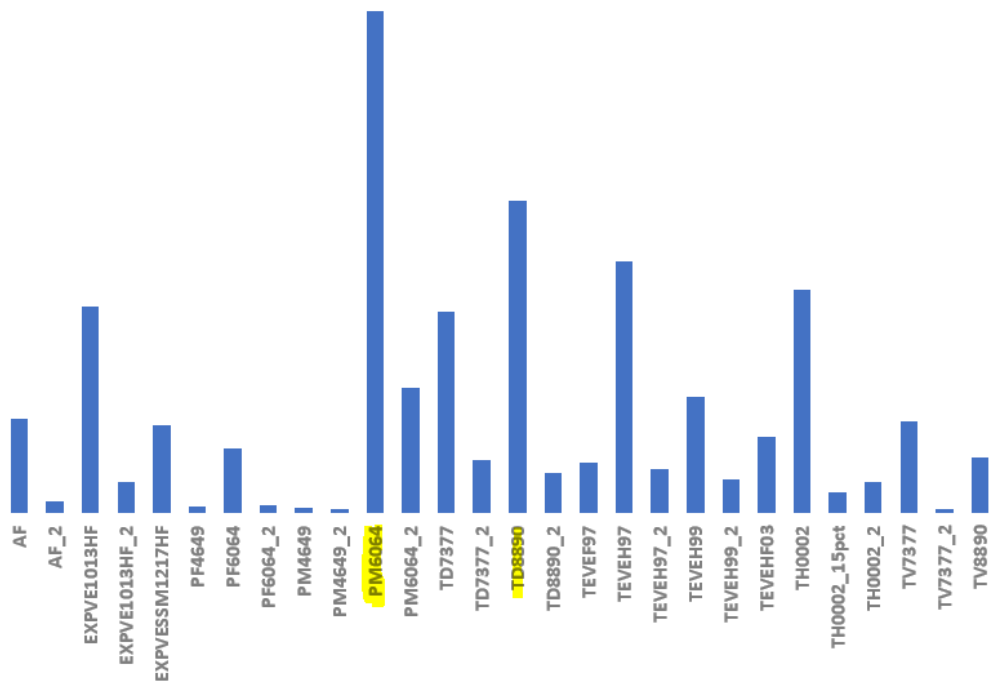
- Code réseau



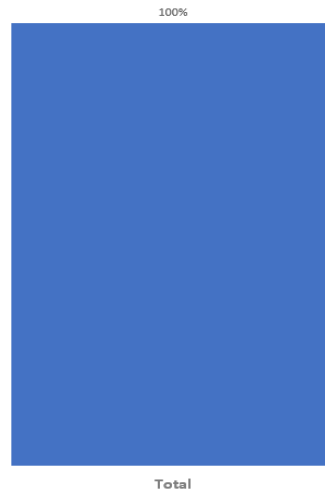
- Famille de produit



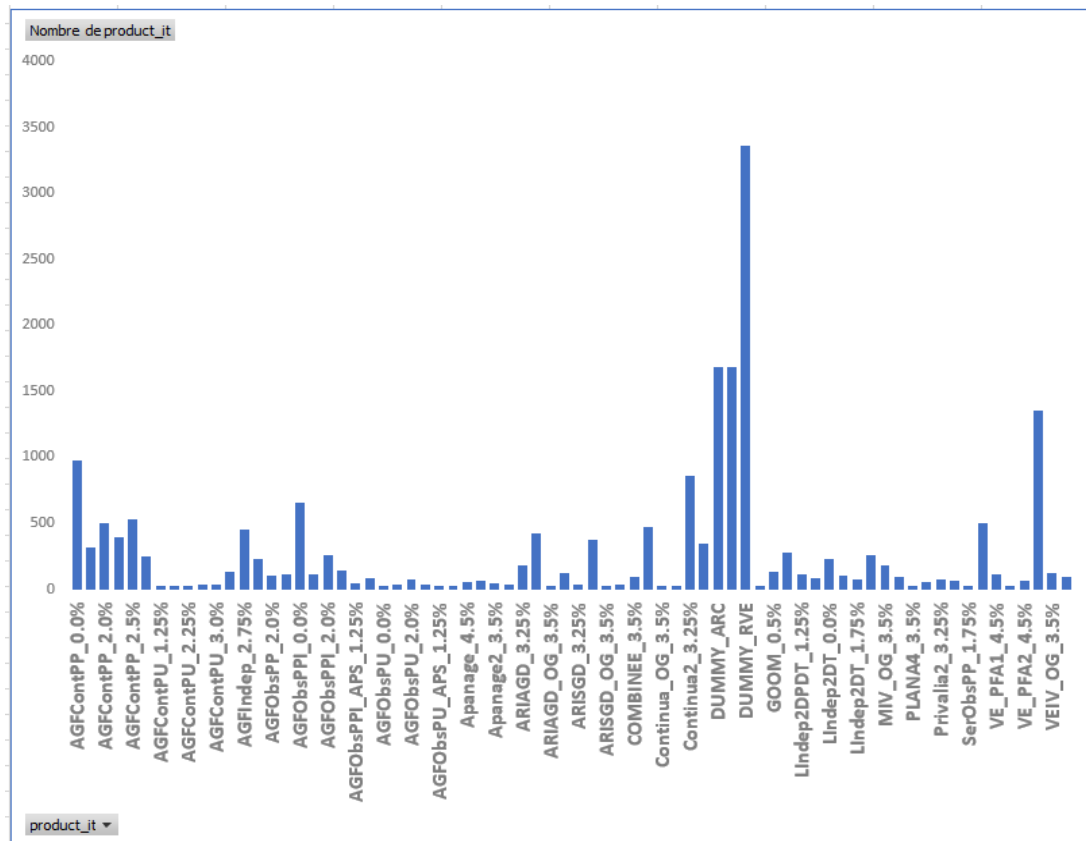
- Famille table de mortalité



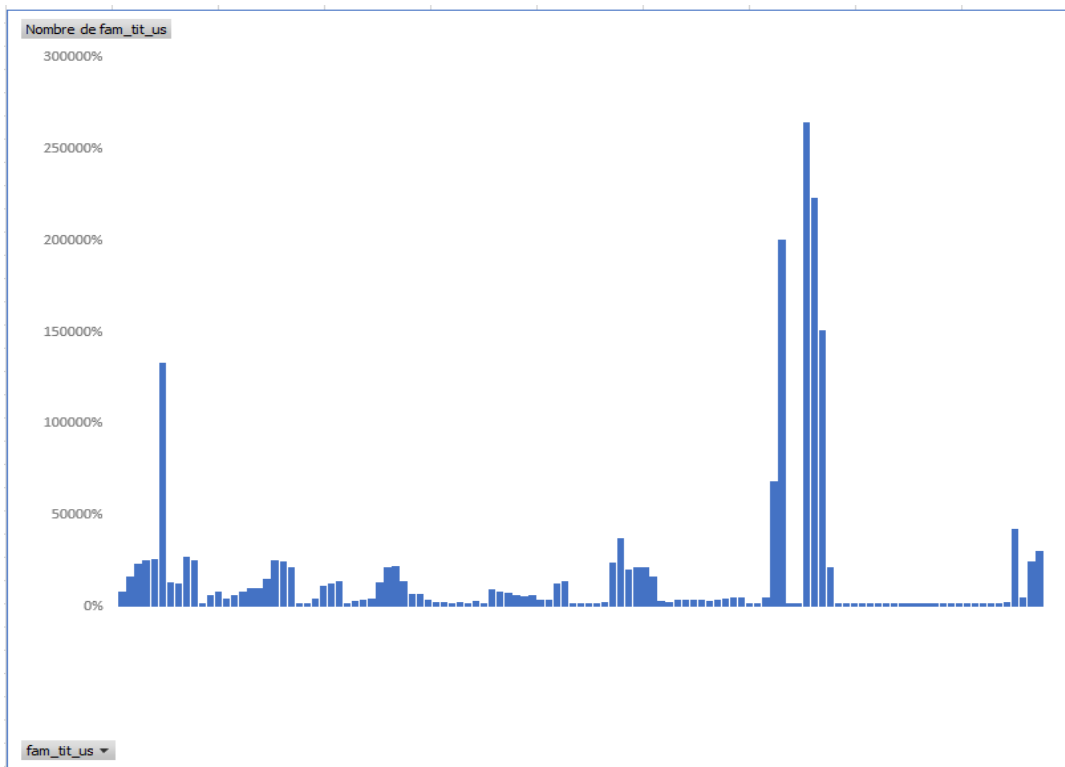
- Fund/Product



- Code produit intérêt technique



- Fam_tit_us



- Variables quantitatives
- Sélection des variables

Variabiles	Nombre de modalités
eval_annee	1
eval_mois	1
prime_pu	1
rescale_x	1
coefcus_ca	2
coeff_zill	2
coefsa_vie	2
indic_majo	2
indic_red	2
freq_pp	4
fam_tit	12
sous_mois	12
gen_usgaap	25
cnt_duree	49
duree_cnt	49
pr_duree	54
duree_pp	55
age_sous	85
sous_annee	92
valid_coef	129
ca_pu_init	244
coefsa_dc	272
coefcus_sa	449
coefcup_sa	545
nb_benefit	604
nb_cnt_ini	604
anc_cnt	891
date_sous	891
tx_chg_acq	6199
prime_pp	8390
sa_pu_init	8597
sa_pp_init	10304
record_no	13524
pm_init	18468
Progressiv	18525

Annexe 2 : Choix de la matrice LOP

- Corrélation au BEL

t	Variable	Corrélation par rapport à la BEL
0	van_pm_clo	0,8707
1	PM d'ouverture (pm_ouvbis)	0,8798
2		0,8874
3		0,9055
4		0,9292
5		0,9463
6		0,9578
7		0,9636
8	Intérêt sur la marge de solvabilite (int_ms)	0,9712
9		0,9797
10		0,9848
11		0,9884
12	Marge de solvabilité (ms)	0,9912
13		0,9929
14		0,9943
15		0,9956
16		0,9962
17		0,9965
18		0,9968
19		0,9971
20		0,9972
21		0,9974
22		0,9975
23		0,9977
24		0,9978
25		0,9980
26		0,9981
27		0,9983
28		0,9983
29		0,9984
30		0,9985
31		0,9985
32	0,9986	
33	0,9987	
34	0,9987	
35	0,9988	
36	0,9989	
37	0,9990	

38		0,9991
39		0,9991
40		0,9992
41		0,9993
42		0,9994
43		0,9995
44		0,9995
45		0,9996
46		0,9996
47		0,9996
48		0,9997
49		0,9997
50		0,9998
51	Cumul des chargements de gestion (chg_ges_cu)	0,9998
52		0,9998
53		0,9998
54		0,9999
55		0,9999
56	Marge de solvabilité (ms)	0,9999
57		0,9999
58		0,9999
59		0,9999
60		0,9999

- Corrélation à la PVFP

t	Variable	Corrélation par rapport à la PVFP
0 . . . 58	EV avant impôt et coût de la marge de solvabilité (fin de mois) (ev_b_f)	1
59	EV après impôt et avant cout de la marge de solvabilité (fin de mois) (ev_n_av_f)	1
60		1

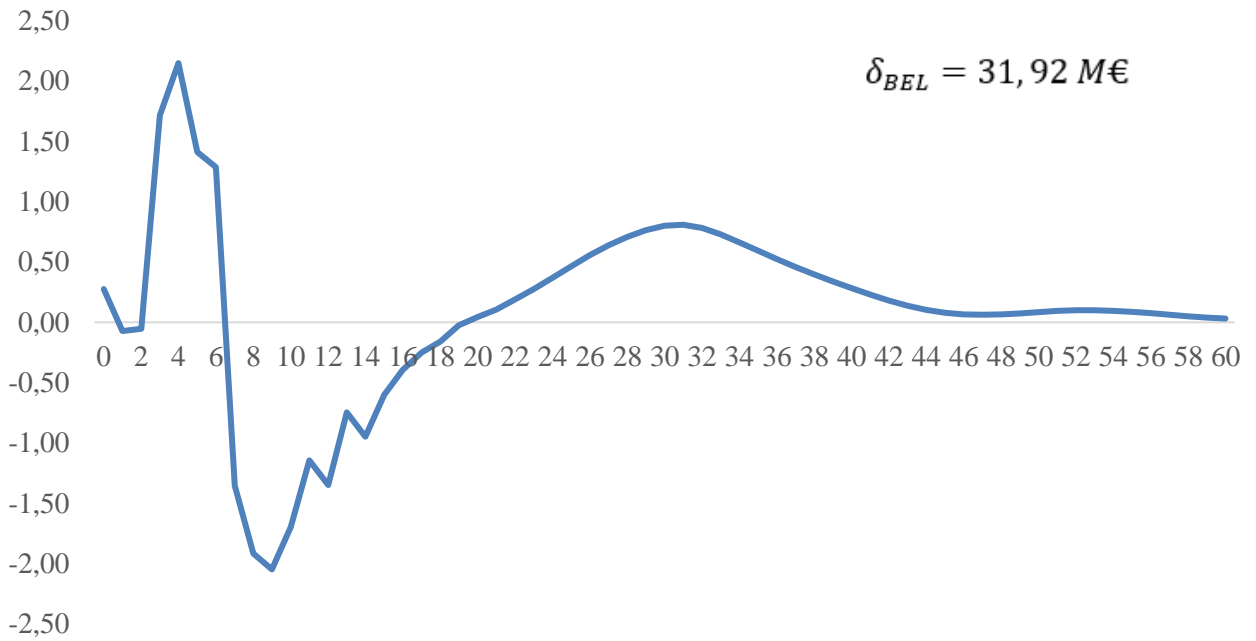
- Corrélation à la BEL et la PVFP

t	Variable	Corrélation par rapport à la BEL	Corrélation par rapport à la PVFP
0	EV net en fin de mois (ev_net_f)	0,5096	0,9993
1		0,5191	0,9990
2		0,4971	0,9985
3		0,4748	0,9986
4		0,4476	0,9984
5		0,4274	0,9983
6		0,4077	0,9984
7		0,3935	0,9978
8		0,3910	0,9976
9		0,3854	0,9974
10		0,3845	0,9971
11		0,3845	0,9967
12		0,3883	0,9961
13		0,3896	0,9973
14		0,3927	0,9969
15		0,3907	0,9969
16		0,3958	0,9968
17		0,3939	0,9966
18		0,3970	0,9955
19		0,4023	0,9941
20		0,4080	0,9937
21		0,4209	0,9932
22	0,4405	0,9942	
23	nb_cnt_f	0,9831	0,4560
24	ev_net_f	0,4382	0,9811
25	Nb de décès du mois (nb_dc)	0,9462	0,4709
26		0,9516	0,4799
27		0,9542	0,4869
28		0,9569	0,4555
29		0,9606	0,4629
30		0,9640	0,4718
31		0,9665	0,4878
32		0,9688	0,4663
33		0,9706	0,4587
34		0,9731	0,4455
35		0,9746	0,4684
36		0,9755	0,4980
37		0,9765	0,4643
38	Nb contrats en cours (fin de mois)	0,9972	0,4466
39		0,9976	0,4487

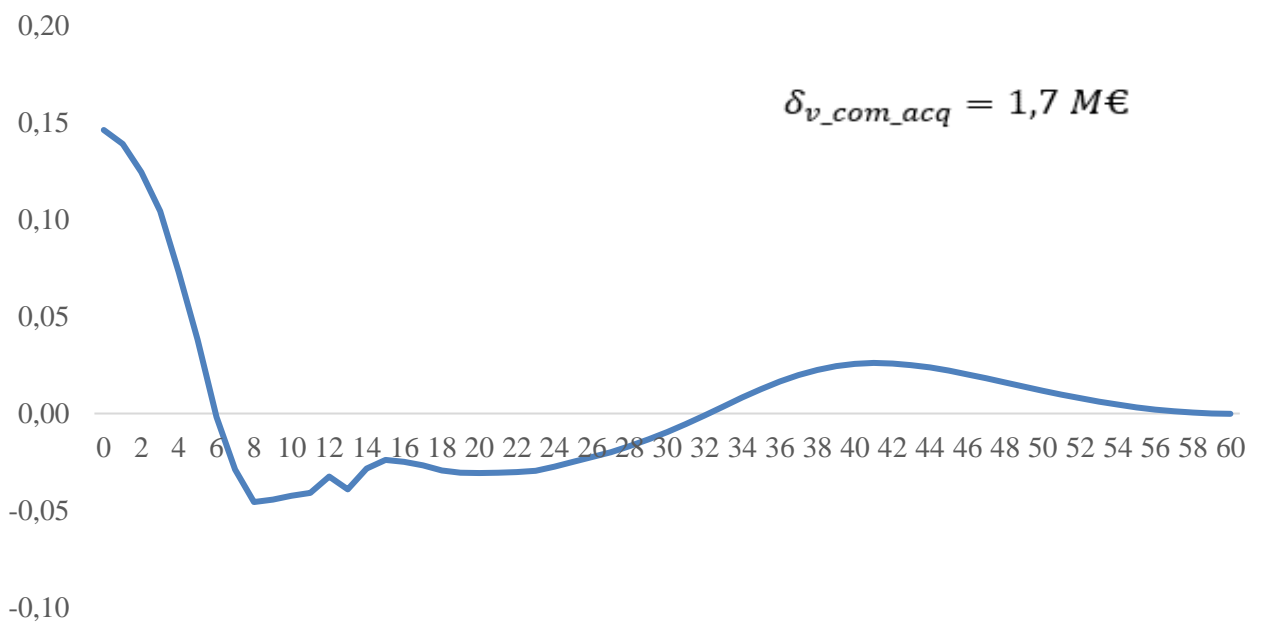
40	(nb_cnt_f)	0,9979	0,4585
41		0,9982	0,4594
42		0,9984	0,4571
43	van_cout_acte_gest (v_act_gest)	0,9075	0,5299
44		0,9076	0,5349
45		0,9069	0,5376
46		0,9066	0,5352
47		0,9062	0,5390
48		0,9058	0,5438
49		0,9057	0,5674
50		0,9053	0,5591
51		0,9034	0,5626
52		0,9000	0,5682
53		0,8970	0,5682
54	Nb contrats en cours (fin de mois) (nb_cnt_f)	0,9998	0,4594
55		0,9998	0,4832
56		0,9998	0,4844
57	ev_net_f	0,5564	0,9523
58	Nb contrats en cours (fin de mois) (nb_cnt_f)	0,9999	0,5831
59		0,9999	0,4635
60		0,9999	0,4510

Annexe 3 : Comparaison des indicateurs constituant le BEL

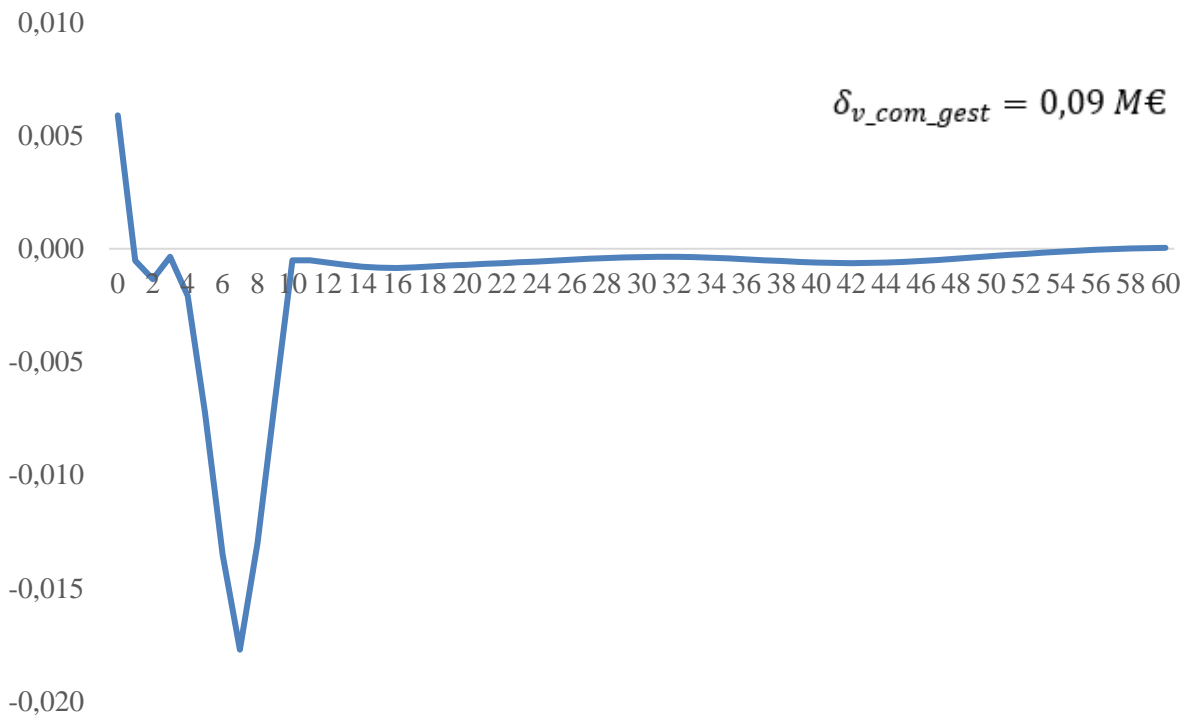
Δ BEL (en M€)
avant agrégation et après agrégation sur la base de la matrice LOP BEL
& PVFP



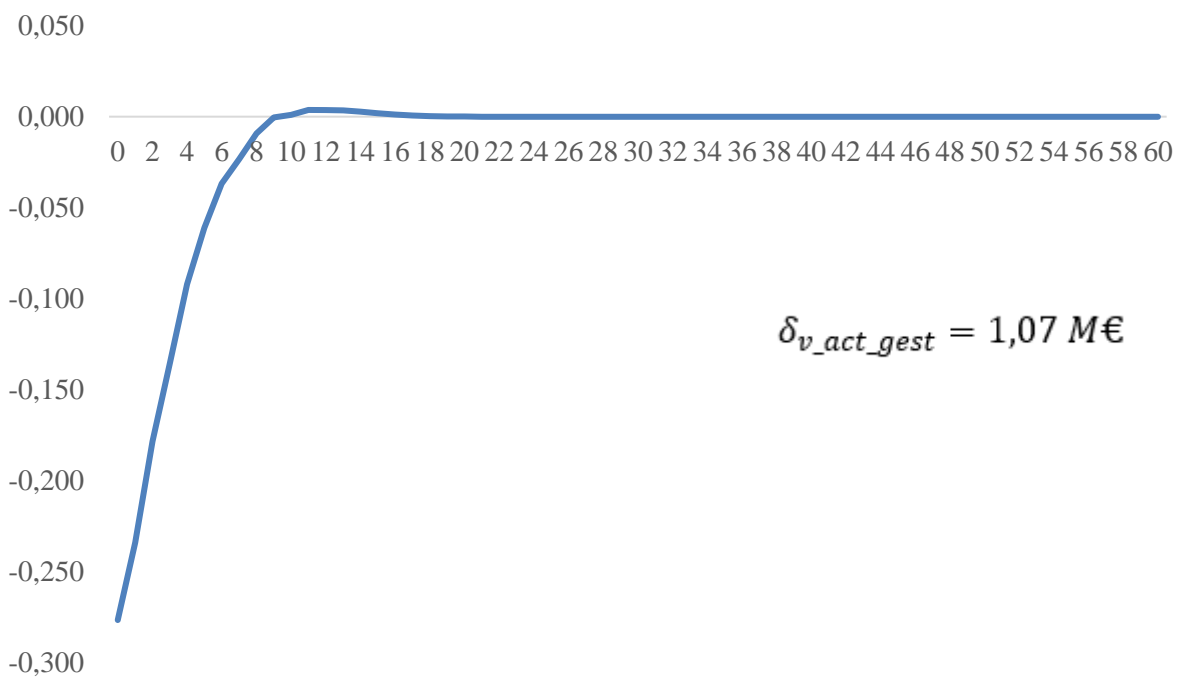
Δ VAN Coût d'acquisition (en M€)



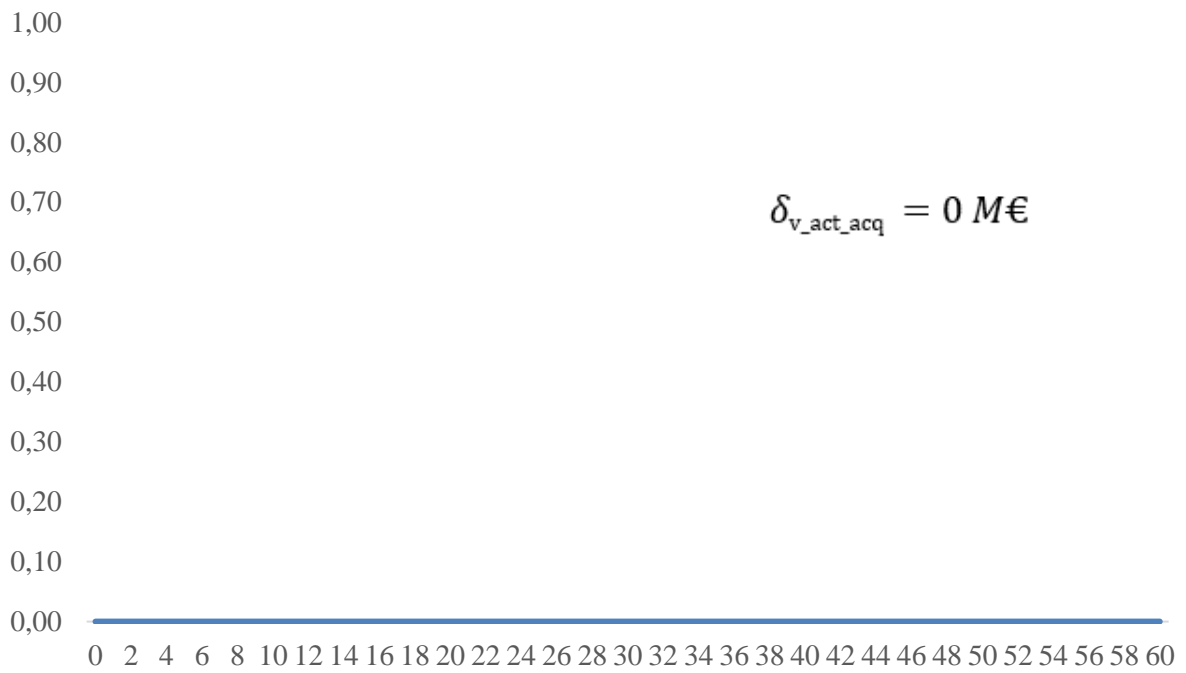
Δ VAN Coût de gestion (en M€)



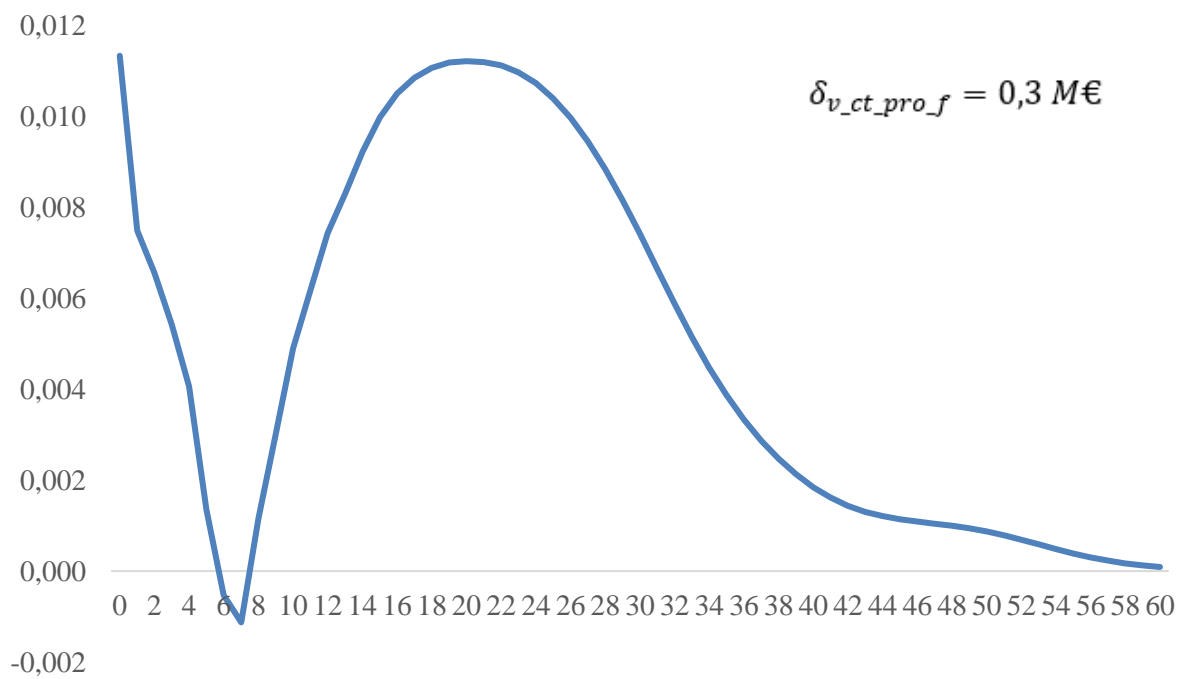
Δ VAN acte de gestion (en M€)



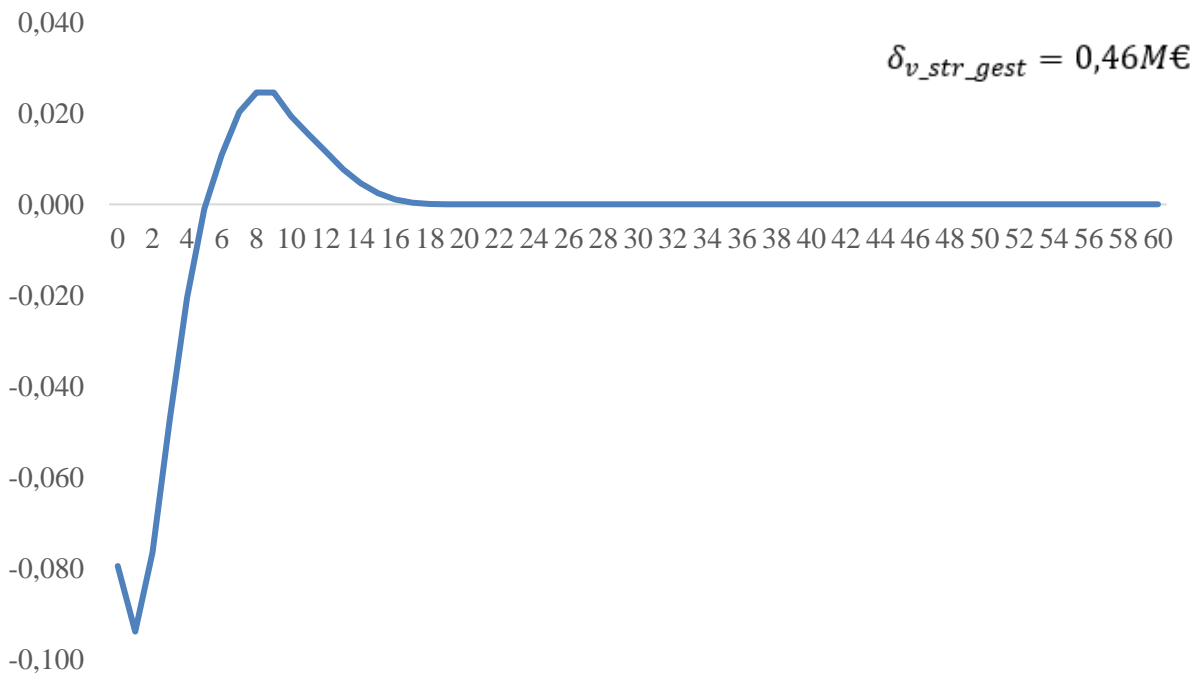
Δ VAN acte d'acquisition (en M€)



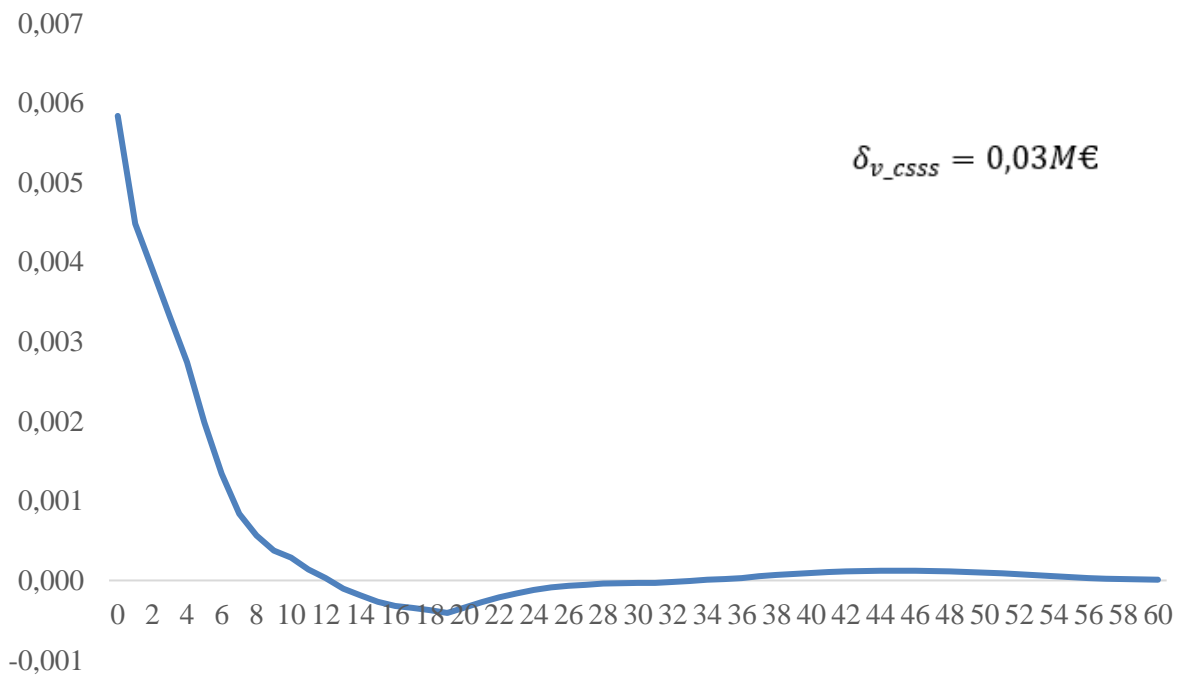
Δ VAN Coût produit financier (en M€)



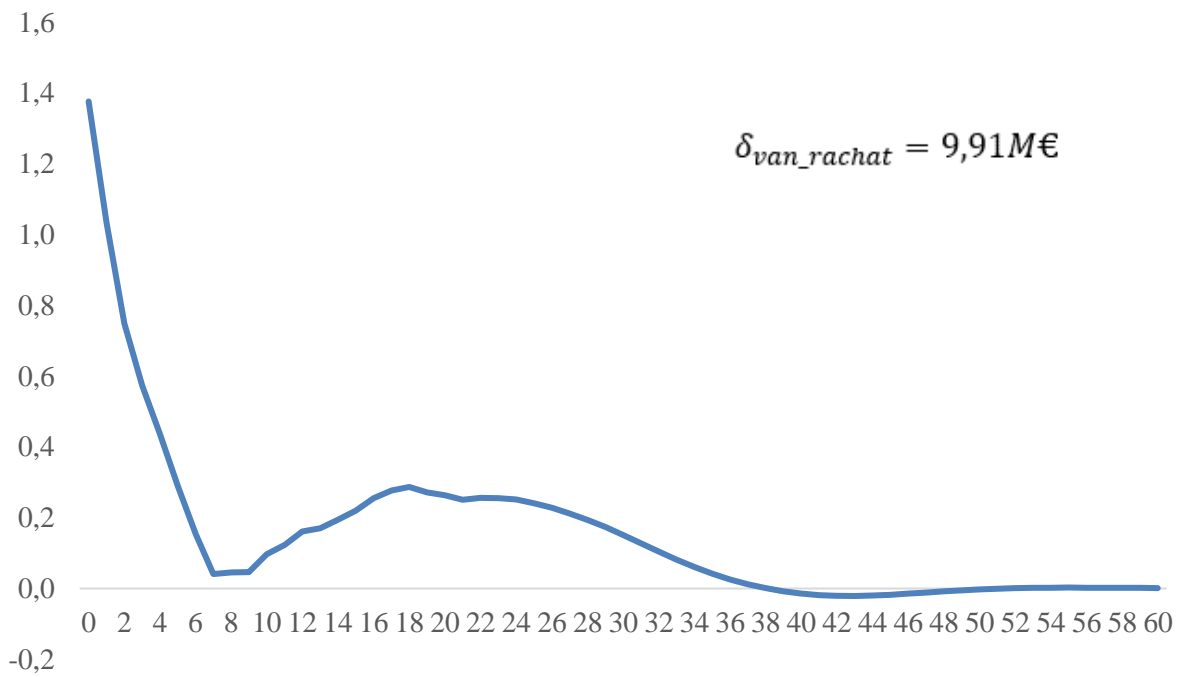
Δ van coût structure de gestion (en M€)



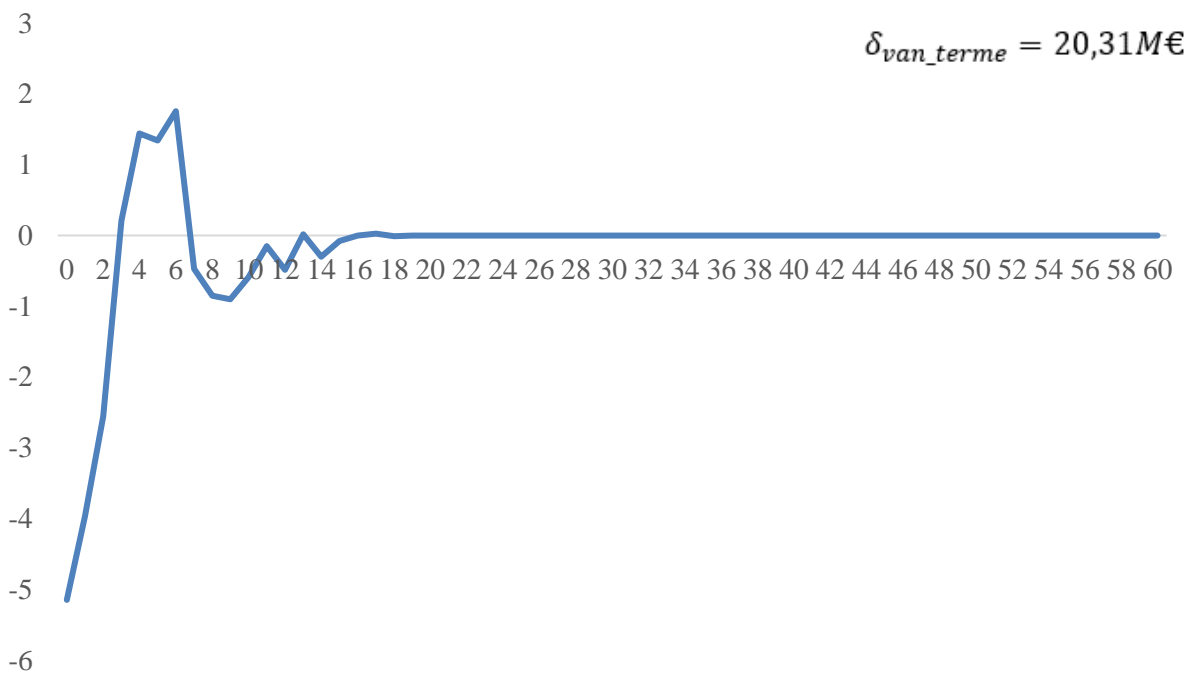
Δ van csss (en M€)



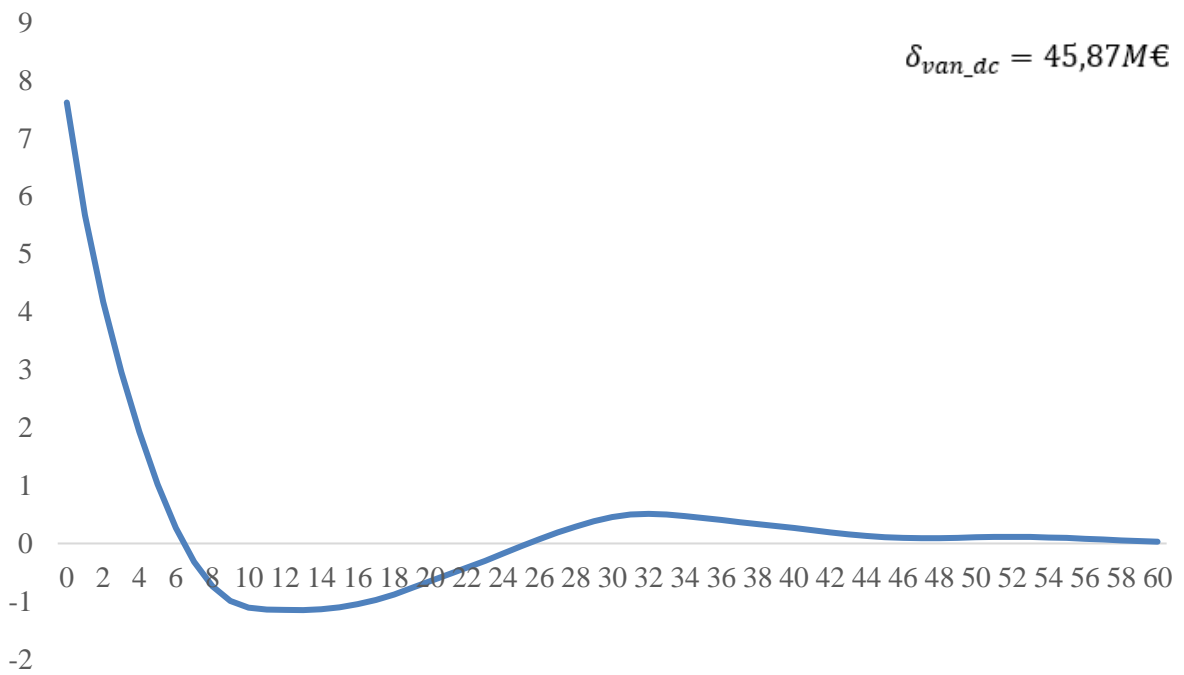
Δ VAN des rachats (en M€)



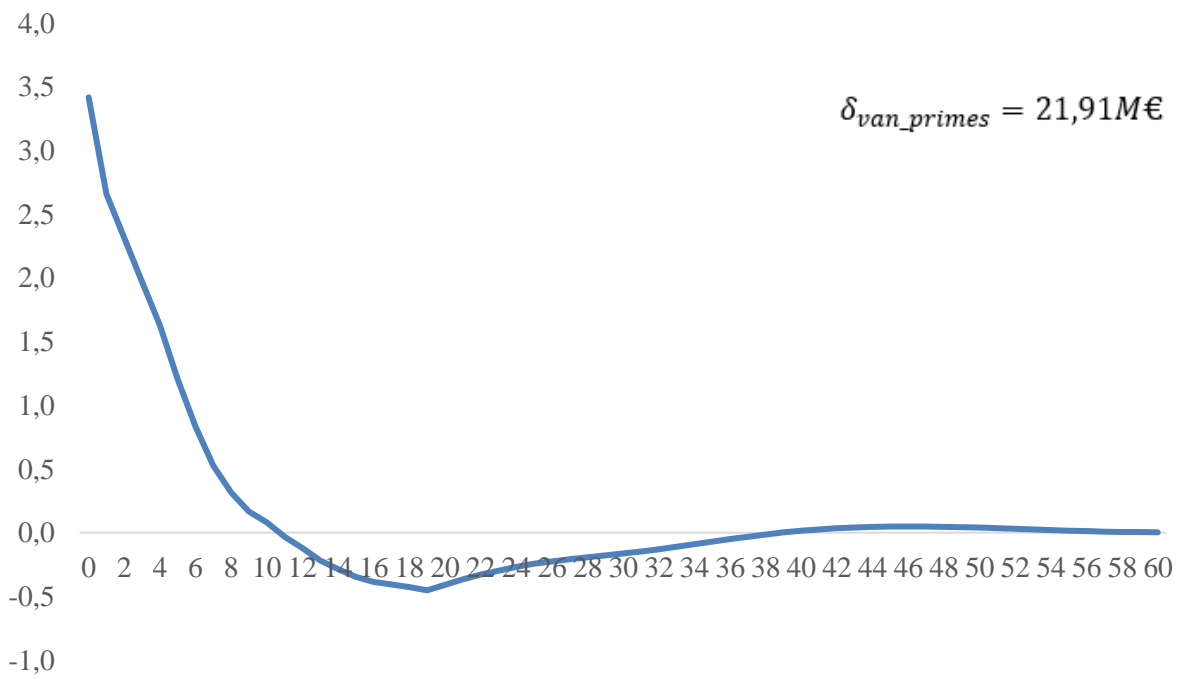
Δ VAN des termes (en M€)



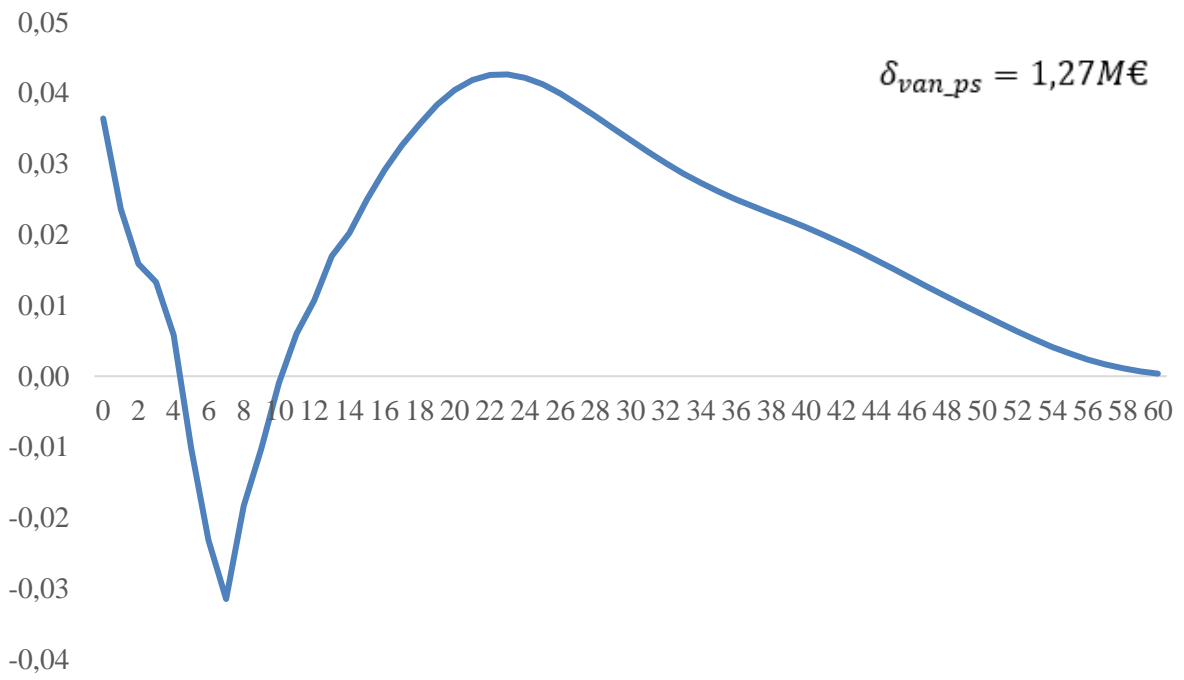
Δ VAN décès (en M€)



Δ VAN primes (en M€)



Δ VAN prélèvement sociaux(en M€)



Annexe 4 : Principe général de l'ACP

L'ACP est une méthode d'analyse factorielle dont les débuts remontent à 1901 avec un article de Karl Pearson. Elle est généralement utilisée pour synthétiser l'information lorsqu'on a une quantité importante de données quantitatives à traiter. Elle permet la transformation de variables initialement corrélées en de nouvelles variables non corrélées.

Chaque nouvelle variable appelée composante principale est une combinaison linéaire des variables initiales. L'ACP est donc une méthode qui permet de condenser l'information en gardant les relations caractéristiques (proximités entre variables et individus).

- Approche de l'ACP et construction des composantes principales

Considérons P variables quantitatives observées sur n individus tel que le tableau de données X suivant :

$$X = \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \dots & & \dots & & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & & \dots & & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}^t$$

Le nuage des points-individus est représenté par les coordonnées des n individus et chaque individu a pour coordonnées :

$$X_i = [x_{i1} \dots x_{ij} \dots x_{ip}]$$

Le nuage des points-variables est représenté par les coordonnées des p variables et chaque variable a pour coordonnées :

$$X_j^t = [x_{1j} \dots x_{ij} \dots x_{nj}]$$

Les représentations du nuage d'individus et du nuage de variables permettent d'analyser les liens entre les individus ou les variables. Malheureusement, il est impossible de faire une telle représentation lorsque la dimension est grande.

Le but de l'ACP et des méthodes factorielles en général est de permettre une telle représentation tout en contrôlant la perte d'informations. L'ACP détermine donc un sous espace E_q de faible dimension ($q \leq p$), sur lequel est projeté le nuage de points du tableau de données de sorte à

conserver les liens significatifs entre ces données. E_q est alors le sous-espace permettant la dispersion (l'inertie) maximale du nuage projeté.

Ce sous-espace est appelé espace factoriel. Les nouveaux axes sur lesquels le nuage est projeté sont appelés les composantes principales.

Soit I l'inertie totale du nuage, G le centre de gravité du nuage et m_i le poids de chaque individu. L'inertie étant la somme des distances des points au centre de gravité pondéré par le poids de chaque point, on a :

$$I = \sum_{i=1}^n m_i d^2(X_i, G)$$

Pour faciliter le calcul d'inertie, il faut transformer les données afin de ramener le centre de gravité à l'origine. De plus, l'inertie étant basée sur la distance, il faut normaliser les données afin d'annihiler les différences des unités ou des échelles de mesures de chaque variable sur la variance. On parle alors dans ce cas d'une ACP normée. Pour la suite, nous allons supposer que les individus ont le même poids et que notre tableau de données est normalisé, donc le centre de gravité est confondu à l'origine.

L'espace factoriel est constitué en déterminant itérativement ces composantes principales. On détermine la première composante principale qui représente la composante pour laquelle l'inertie projetée est la plus grande. Les autres composantes sont obtenues par la même approche. Considérons $\overrightarrow{Gu_1}$ le vecteur directeur de la première composante principale et I_1 l'inertie projetée sur cette composante, il faut :

$$I_1 \text{ maximale}$$

$$\text{avec } \|\overrightarrow{u_1}\| = 1$$

Soit c_{i1} la projection orthogonale de l'individu X_i sur la première composante principale :

$$c_{i1} = \overrightarrow{X_i} \cdot \overrightarrow{u_1}$$

L'inertie projetée sur la première composante est alors :

$$I_1 = \sum_{i=1}^n c_{i1}^2$$

Soit C_1 le vecteur de projection des n individus sur cette composante :

$$C_1^t = [C_{11} \dots C_{ij} \dots C_{nj}]$$

On a donc :

$$C_1 = X \cdot u_1$$

On en déduit que :

$$I_1 = C_1^t \cdot C_1 = u_1^t X^t X u_1$$

Soit la matrice d'inertie $S = X^t X$, l'inertie projetée peut se mettre sous la forme :

$$I_1 = u_1^t S u_1$$

Il faut donc choisir u_1 de sorte à maximiser $u_1^t S u_1$. On utilise la méthode des multiplicateurs de Lagrange pour déterminer u_1 . On pose donc :

$$L(u_1) = u_1^t S u_1 - \lambda_1 (u_1^t u_1 - 1)$$

La dérivée matricielle donne :

$$\frac{\delta L(u_1)}{\delta u_1} = 2S u_1 - 2\lambda_1 u_1$$

On a alors :

$$S u_1 - \lambda_1 u_1 = 0$$

$$u_1^t S u_1 - u_1^t \lambda_1 u_1 = 0$$

Sachant que $\|\vec{u}_1\| = 1$, on en déduit :

$$u_1^t S u_1 = u_1^t u_1 \lambda_1 = \lambda_1$$

D'où

$$I_1 = \lambda_1$$

λ_1 est la valeur propre la plus grande de S .

On détermine ensuite la deuxième composante sur laquelle le nuage se déforme le moins en projection, après la première composante, tout en étant orthogonal au premier. On réitère jusqu'à l'obtention de p composantes. L'inertie de la k ème composante est obtenue par la même approche et l'inertie projetée sur la k ème composante est :

$$I_k = \lambda_k$$

Les composantes ainsi déterminées étant toutes orthogonales deux à deux, on a donc :

$$I = \sum_{k=1}^p I_k = \sum_{k=1}^p \lambda_k$$

Dans l'espace factoriel ainsi déterminé, les axes sont non corrélés et ne véhiculent pas la même information selon leur rang : leur capacité à « résumer » le nuage se détériore au fur et à mesure que l'on observe des axes de rang élevé

BIBLIOGRAPHIE

Begliomini Sandy [2019], *Alternatives aux portefeuilles répliquants pour le calcul des fonds propres économiques*, Mémoire d'actuariat, ENSEA

COMMISSION REGULATION (EU) 2021/2036 [2021], *amending Regulation (EC) No 1126/2008 adopting certain international accounting standards in accordance with Regulation (EC) No 1606/2002 of the European Parliament and of the Council as regards International Financial Reporting Standard 17*, Official Journal of the European Union.

Donghui CHEN et Robert J. PLEMMONS [2009], *Nonnegativity Constraints in Numerical*, Wake Forest University, Winston-Salem.

Ding C & He X [2004], *K-means Clustering via Principal Component Analysis*, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

Félix WELENC [2013], *Modèle interne partiel en prévoyance individuelle*, Mémoire d'actuariat, CEA.

IASB [2017], *IFRS 17 Insurance Contracts*.

IASB [2017], *IFRS 17 Insurance Contracts, Basis for conclusion*.

Miguel Cuevas UROSA [2014], *Optimization of Model Points*, Master's Thesis in Financial Mathematics, University of Tartu Faculty of Mathematics and Computer Science.

Mingoti Sueli A. et Lima Joab O [2005]. « *Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms* ». In ScienceDirect

Nexialog Consulting [2021], *IFRS 17 Risk Adjustment, Une étude comparative sur le périmètre des contrats d'épargne en euros*, Paris, France.

Stéphane TUFFERY [2008], *Data Mining et Statistique décisionnelle*, Technip

Reda AIT M'BARK [2017], *Approche d'agrégation de contrats d'assurance sous IFRS 17*, Mémoire d'actuariat, ISFA.

Renaud BAILLY et Nicolas GEMIN [2018], *IFRS 17 : interprétation de la norme, premiers résultats et leviers de pilotage pour un portefeuille dommages*, Mémoire d'actuariat, CEA.

Ricco RAKOTOMALALA, *Analyse en Composantes Principales (ACP)*, Université Lumière Lyon 2.

Zha, H., Ding, C., Gu, M., He, X., & Simon, H. [2002]. *Spectral relaxation for K-means clustering*. Advances in Neural Information Processing Systems 14 (NIPS'01), 1057–1064.

LISTE DES FIGURES

Figure 1 : les différentes étapes de la norme IFRS 17 depuis sa publication.....	3
Figure 2 : Comparaison bilans IFRS 4 vs IFRS 17	4
Figure 3 : Détermination de la CSM	7
Figure 4 : Structure d'un modèle déterministe.....	14
Figure 5 : Processus <i>flexing</i>	15
Figure 6 : Modules du modèle ALM.....	16
Figure 7 : Architecture du modèle ALM.....	19
Figure 8 : Processus <i>Full ALM</i>	20
Figure 9 : Description des variables qualitatives	27
Figure 10 : Distribution du portefeuille par produit en fonction du nombre de MP	28
Figure 11 : Répartition des capitaux assurés en fonction du nombre de MP	29
Figure 12 : Répartition des MP en fonction de l'âge de souscription des assurés.	30
Figure 13 : Diagramme en moustache de l'âge de souscription.....	30
Figure 14 : Distribution des MP en fonction des taux techniques	31
Figure 15 : Matrice de corrélation des variables quantitatives.....	32
Figure 16 : Construction de la seconde base de données	34
Figure 17: Contribution des MP au BEL initial	35
Figure 18 : Processus LOP	46
Figure 19 : Calibrage dans le processus LOP	49
Figure 20 : Inertie intra-classe vs inertie inter-classes	58
Figure 21 : Classification K-Means	59
Figure 22 : Nombre de clusters en fonction du coût du modèle (BEL, PM et PVFP)	62
Figure 23 : Nuage de points en trois dimensions	62
Figure 24 : Nombre de facteurs en fonction de la variance expliquée.....	68
Figure 25 : Eboulis des valeurs propres	69
Figure 26 : Cercle de corrélation axes 1 et 2.....	70
Figure 27 : Diagramme à moustache de la variable « sous_annee » (année de souscription) .	72
Figure 28 : Diagramme à moustache de la variable «sa_pp_init »	72
Figure 29 : Cercle de corrélation axes 3 et 4.....	74
Figure 30 : Nombre de clusters- K-Means sur ACP	76

LISTE DES TABLEAUX

Tableau 1 : Description quantitative du portefeuille	25
Tableau 2 : Description des variables quantitatives	26
Tableau 3: Contribution des différentes garanties au BEL	35
Tableau 4 : Variables corrélées au BEL	51
Tableau 5 : Variables corrélées à la PVFP	52
Tableau 6 : Variables corrélées au BEL et la PVFP	53
Tableau 7 : Synthèse des vecteurs LOP par critères et impacts BEL et PVFP	55
Tableau 8 : Clusters basés sur les variables run MP / MP	63
Tableau 9 : Agrégation sur la base du Clustering K-Means	64
Tableau 10 : Matrice de corrélation	66
Tableau 11 : Valeurs propres et variances expliquées	67
Tableau 12 : Représentation des variables sur le premier axe factoriel	71
Tableau 13 : Représentation des variables sur le second axe factoriel	73
Tableau 14 : Représentation des variables sur le troisième axe factoriel.....	75
Tableau 15 : Représentation des variables sur le quatrième axe factoriel.....	75
Tableau 16 : Clusters basés sur ACP sur les MP	76
Tableau 17 : Agrégation sur la base du Clustering K-Means après ACP	77